Aladdin, Alla dien, Allendien, please evaluate the performance of Whisper on Dutch dysarthric speech

Thesis submitted for the MSc Voice Technology

Lian Feenstra (s4957180)

Supervisor: Dr. Shekhar Nayak

Second reader: Dr. Matt Coler

August 19, 2023

rijksuniversiteit
groningen

campus fryslân

1 Abstract

Automatic Speech Recognition (ASR) will take spoken speech and convert this into text. Although this technology is now readily available to most of us, as of now this technology is not usable for everyone. People with dysarthria experience a worse performance in ASR compared people without dysarthria. This is mostly due to the smaller amount of dysarthric data available and the characteristics present in dysarthric speech. The severity of the dysarthria impacts the ASR performance for that speaker. Recently the Whisper language processor was released. The creators of Whisper propagate the processor as a robust system. In this research the weakly supervised encoder-decoder Transducer (Whisper) architecture is compared to a hybrid model in a Dutch dysarthric speech recognition task. The Whisper model is finetuned on dysarthric data from the COPAS corpus and evaluated on the Domotica-3 dataset. The weakly supervised encoder-decoder Transducer did not manage to outperform the hybrid model. Results showed a limited decrease in the WER for speakers with moderate and high severity dysarthria. The speakers with mild severity dysarthria suffered worse performance after finetuning. The weakly supervised encoder-decoder Transducer did not outperform the hybrid ASR model on any of the speaker used in the evaluation. Future research is needed to make ASR systems truly accessible to the demographic of dysarthric speakers. This research could focus on evaluating the Whisper architecture using a different pretraining strategy, using transfer learning or using a different evaluation dataset. Another possibility would be to evaluate the speech of dysarthric speakers and get a clearer view of of which characteristic of the speech of dysarthric speakers seems to degrade the ASR performance.

Contents

2 Introduction

Speech allows us to communicate to each other and to voice our thoughts and needs. We can use speech to tell our friends about our weekend, tell our dogs to sit and in recent years, to use our computer of phone. For us to be able to speak to our computers or phones, they need to understand what we are saying.

Automatic Speech Recognition (ASR) will take spoken speech and convert this into text. An ASR system can be constructed in a number of different ways. Once a system is developed it needs to be trained. These systems are trained on a large amount of speech data to learn to predict the right transcriptions for new speech data. ASR is used in our daily lives, for example in the form of virtual assistants and/or in home automation. A virtual assistant uses an ASR system combined with a system that can understand or interpret the spoken text. Speech recognition and its implementations can help people in their daily lives. The Dutch government advises the use of virtual assistants to help elderly people, or people of dementia or Parkinson's disease to live at home longer on one of their websites [1]. Speech recognition is also making its way towards health care and speech therapy. It can be used in aphasia therapy, for people suffering from aphasia to practise at home [2]. There are also initiatives for the use of ASR in rehabilitation care [3].

With ASR being used in our homes and in our health care system, this technology should be available for everyone to use. Unfortunately this isn't the case yet, as ASR systems do not perform as well for everyone. The data present in the training data is important in the performance of the ASR. Research on Dutch speech recognition showed worse ASR performance on accented speech, children and elderly voices (Feng, Kudina, Halpern, & Scharenborg, 2021). When compared to typical speech, ASR systems also perform worse on speech of people with dysarthria (Moore, Venkateswara, & Panchanathan, 2018).

Although ASR is already used in many settings, including healthcare, these systems are not accessible to every user. For this reason it is important to research how ASR can be improved for the group of speakers with dysarthria. There are multiple ways to construct an ASR system. As different models are constructed in different ways, they perform differently on dysarthric speech. The goal of this research is to explore which of two ASR architecture can perform best on dysarthric speech. To achieve this goal, a weakly supervised encoder-decoder Tranducer and a hybrid model are compared. The models are the Whisper architecture (Radford et al., 2022) and a TDNN acoustic model architecture as reported by Wang, BabaAli, et al. (2021). First a theoretical background is given, here dysarthria, the problems dysarthric speech faces in ASR, ASR in general and the Whisper architecture are elaborated/discussed. The results for the different ASR models are compared using the Word Error Rate (WER). Different finetuning setups are explored to find the best performance on a Dutch dysarthric speech recognition task.

For this thesis the choice was made to talk about 'typical speech' when talking about the speech of people without dysarthria, and 'speech of people with dysarthria' or just 'people with dysarthria'. This is different from most papers where the term 'normal' or 'healthy' speech is used. This choice was made to not only reduce a person to their speech disorder but to also talk about the person behind this disorder. Sometimes, when not talking about a specific speaker, the term 'dysarthric speech' is used.

[1] The Dutch government advises the use of virtual assistants to help people live independent longer https://www.zorgvannu.nl/innovaties/langer-zelfstandig-thuis-met-een-spraakassistent

[2] Afasietherapie, an app created for people with aphasia, to be used for aphasia therapy, that uses speech recognition https://apps.apple.com/nl/app/afasietherapie/id1459635403.

[3] Virtual reality applications are developed for rehabilitation therapy. Speech recognition is used in the speech therapy applications. https://www.vogellanden.nl/actueel/symposium-bij-vogellanden-over-vr-in-de-revalidatiezorg

3 Dutch dysarthric speech recognition

3.1 Dysarthria

Clear communication does not come naturally to everyone. A person might struggle with the pronunciation of speech after for example a stroke or when suffering from Parkinson's disease, this is called dysarthria.

There are a lot of differently phrased definitions of dysarthria available. Dysarthria is usually described a motor speech disorder (De Russis & Corno, 2019; Jaddoh, Loizides, & Rana, 2022) or a sensorimotor speech disorder (Lowit & Kent, 2016, p.529).

Duffy (2019, p.3) defines dysarthria as 'a collective name for a group of neurological speech disorders that reflect abnormalities in the strength, speed, range, steadiness, tone or accuracy of movements required for breathing, phonatory, respiratory, articulatory, or prosodic aspects of speech production' This is the definition that will be followed during this research. As this definition gives a good overview of the different aspects that can be affected by dysarthria.

Depending on the person with dysarthria, different characteristics of dysarthric speech can be differentiated. These characterics include, but are not limited to, flaccid articulation, hyper or hypo nasality, hoarse voice, change in intonation (slowed, monotonous or too fast), distorted vowels and inaccurate pronunciation. The errors a person makes during speaking are consistent (Sluijmmers, Singer, Versteegde, Zoutenbier, & Gerrits, 2016).

These characteristics can occur in different levels of severity. There are multiple ways available to test the severity levels and the type of dysarthria. One way to measure the severity of the dysarthria is by measuring the intelligibility of the speech. The intelligibility of a speaker shows us how well a spoken signal is understood. This is different from communication effectiveness, which also takes into account non-verbal cues and the context of the message (Lowit & Kent, 2016, p.535) In a Dutch speech therapy setting, the speech intelligibility can be measured using the Dutch Intelligibility Assessment (De Bodt, Guns, & Van Nuffelen, 2006). This test will give an intelligibility score. The intelligibility score can be used to measure the severity of the dysarthria and to measure the effectiveness of the speech therapy intervention. This can be done by measuring the intelligibility score before and after a therapeutic intervention and comparing these scores. In the NSVO-Z, a sentence task of the NSVO, the person with dysarthria has to read illogical sentences without the speech therapist seeing these sentences beforehand. After transcribing these sentences and comparing them to the target sentences, a score can be reported. This is a good example of measuring intelligibility and not communication effectiveness, as there are no contextual cues present.

Dysarthria is often classified in multiple different categories based on lesion site and common characteristics of the speech. During the diagnostics of dysarthria, the speech of a person with dysarthria can be categorised as flaccid dysarthria, spastic dysarthria, ataxic dysarthria, hyperkinetic dysarthria or hypokinetic dysarthria (Darley, Aronson, & Brown, 1969). These types of dysarthria can also occur together. For example, the speech of people with Amyotrophic Lateral Sclerosis (ALS) is characterized as a combination of flaccid and spastic dysarthria (Lowit & Kent, 2016). The cause of dysarthria can be congenital conditions (e.g. cerebral palsy), neurologic injury (e.g. brain trauma and/or strokes) or diseases (e.g. neuromuscular diseases, Parkinson's disease, ALS, Huntington's disease ect.) (Lowit & Kent, 2016).

For this research it is important to know there are different types of dysarthria with different characteristics. There is a large variability in dysarthric speech. This could be of importance in researching dysarthric speech recognition.

The exact numbers on the prevalence of dysarthria in the Netherlands are not available. It is estimated that around 15% of people in the chronic phase after a cardiovascular accident suffer from dysarthria, for people with Parkinson's disease the prevalence of dysarthria is around 70%. In the group of neuromuscular diseases the prevalence is between 42% and 62% (Sluijmmers et al., 2016). Although these percentages don't give exact numbers, they do give an indication of the amount of people affected by this speech disorder.

Dysarthria should not be looked at in only a medical point of view. As it can have a big impact on a persons quality of life and limits people in their daily participation in society. Dysarthria lead to isolation and fear of communication, this frequently leads to speakers avoiding certain speaking contexts (Lowit & Kent, 2016). Dysarthria can lead to isolation, a negative self image and depression. This can result in avoidence of certain speaking contexts (Lowit & Kent, 2016; Sluijmmers et al., 2016). Not only can people with dysarthria struggle while speaking to other people, or even avoid these situations, their interactions with ASR systems can also be influenced by their dysarthria.

3.2 The difficulties of ASR in dysarthric speech

When comparing the performance of typical and dysarthric speech in current ASR systems, the performance is worse on dysarthric speech compared to the speech of people without dysarthria (Moore et al., 2018). Multiple causes for this difference in performance can be identified. One limitation in the creation of speech recognition systems suited for dysarthric speech recognition, is the limited availability of dysarthric/pathological speech databases. The databases that are available are often much smaller compared to databases for typical speech (Yılmaz, Mitra, Sivaraman, & Franco, 2019). This means less training data is available to train the ASR systems. This problem is enhanced due to the large variability in speech present in the speech of people with dysarthria (Moore et al., 2018).

The accuracy of an ASR system is also influenced by the severity and the intelligibility of the person with dysarthria (Mustafa, Rosdi, Salim, & Mughal, 2015). Lower intelligibility leads to worse ASR performance.

As mentioned above, one of the characteristics of dysarthric speech is the abnormality in speed (Duffy, 2019). People with certain types of dysarthria speak more slowly and/or there is segmentation of syllables when speaking. Many ASR systems can interpret this slower speaking rate and the segmentation of syllables as the start of a new word, this leads to a worse performance of the ASR system (Moore et al., 2018; Young & Mihailidis, 2010).

Both Jaddoh et al. (2022) and Young and Mihailidis (2010) found similar trends with regards to difference in performance in commercial ASR systems between typical and dysarthric speech. Even though these literature reviews were performed twelve years apart. Young and Mihailidis (2010) even mentions literature from early 1990 to early 2000 showing the same trends. These studies show a decrease in ASR performance with increased severity of dysarthria. This shows that speech recognition of dysarthric speech is still performing worse than typical speech, despite improvements in ASR technology.

The lesser performance of ASR systems on dysarthric speech, means this type of technology is less accessible to the demographic of people with dysarthria. More research is needed to develop an ASR system that performs well on dysarthric speech. Other than accessibility, an ASR system with good performance could also be used in a speech therapy setting. An example would be for the assessment of the dysarthria or for feedback for people to practise at home if they have no one to practise with. A good performing ASR system could also help the development of communication devices, helping people who struggle to communicate clearly to be understood by the people around them and regain some confidence and independence that can be lost with dysarthria.

3.3 Automatic Speech Recognition

Automatic Speech Recognition (ASR) takes spoken language and converts this to written text [4]. The first ASR research was conducted in 1952 and resulted in a system that could recognize single digits, spoken by a single speaker. This system was created by Bell Laboratories. Trough out the 1960s and 1970s these developments continued towards an ASR system that could recognize 1011 different words (Kamath, Liu, & Whitaker, 2019).

Around the 1980s HMMs, a statistical/probabilistic approach to speech recognition, became a dominant technique for ASR, around this time neural networks were also introduced in speech recognition. In the 1990s one of the commercially available software was 'Dragon', which could recognise around 80 000 words (Kamath et al., 2019).

There are different methods to constructing an ASR model. One type of ASR model is called a hybrid ASR system. This contains an acoustic, pronunciation and language model to achieve the recognition of spoken language. In hybrid ASR systems the acoustic model predicts a likelihood of the acoustic input speech given a phoneme sequence (Rao, Sak, & Prabhavalkar, 2017). This means the model takes a waveform as its input and will calculate how likely a certain phoneme sequence is pronounced. The pronunciation model is a dictionary of the pronunciation of words, often made by linguists, also containing solutions for words not present in the dictionary (Rao et al., 2017). With this pronunciation model, the ASR system knows how different words can be pronounced, keeping in mind different variations a word can be pronounced. The language model can be a N-gram model trained on text data in the target language (Rao et al., 2017). This means the model calculates the probability of words occurring together in the string.

Another type of ASR model is an end-to-end (E2E) model. In an E2E model, the model directly outputs a transcription given an audio as input (Rao et al., 2017). The components present in a traditional ASR system, the acoustic, pronunciation and language model, are replaced with a single model. This can make the model size smaller, and make the models more attractive for use on, for example, mobile devices (He et al., 2019; B. Li et al., 2020). In common benchmark datasets, E2E models receive state of the art results (J. Li et al., 2022; Vielzeuf & Antipov, 2021).

There are multiple approaches for machine learning that can be used in ASR development. Although there are more possibilities available, the methods that are relevant for this study, are mentioned here. In *supervised learning* each example in the dataset is labeled. The model uses this labeled data to predict output. This can be used for classification or regression (Kamath et al., 2019).

In *unsupervised learning* there are no labels available in the dataset. The model will cluster this data based on similarity. An example could be a system that makes recommendations for a user based on early watched content (Kamath et al., 2019).

Another option is *weakly supervised learning*. Here the dataset is labeled, but these labels are of lower quality. This might mean the labels are not produced by experts in a certain field.

An ASR model can be personalised or non-personalised. In a personalised ASR system the model is finetuned on the data of a specific speaker. This can help the performance of the ASR system for that specific speaker. In non-personalised ASR systems this isn't the case. These systems aren't trained or finetuned on a specific speaker. During this thesis the focus will be on non-personalised ASR.

[4] Essential Guide to Automatic Speech Recognition Technology. `https://developer.nvidia.com/blog/essential-guide-to-automatic-speech-recognition-technology/`

3.4 Word Error Rate

A common way to measure the performance of ASR is the Word Error Rate (WER) (Kamath et al., 2019). The WER is calculated by comparing the reference transcription to the hypothesised transcription made by the ASR system. The calculation for the WER is presented in figure 1 (Kamath et al., 2019). This is calculated by dividing the amount of substitutions (S), insertions (I) and deletions (D) made in the output transcription by the total amount of words spoken in the correct transcription (N). These are the type of errors often encountered in the output of ASR systems. The end result of the WER calculation is a percentage that can be used for measuring the performance of the ASR system. This also gives the possibility to compare different ASR systems, or iterations of ASR systems, to one another.

Figure 1: WER calculation

$$WER = 100 \times \frac{I + D + S}{N}$$

3.5 Whisper ASR

Whisper is a language processor system first introduced by Radford et al. (2022). Whisper uses an encoder-decoder Transformer as their architecture, this architecture was first proposed by Vaswani et al. (2017). The encoder-decoder Transformer is a type of E2E model. The Whisper architecture is visualised in figure 2 (Radford et al., 2022).

The language processor system contains different possibilities. The processor is a multilingual model. Whisper has multilingual training data for 75 different languages, there is a correlation between the amount of training data and the WER. The authors estimate that a 16 times increase in training data for a specific language, will reduce the WER by half (Radford et al., 2022).

The language processor is not only able to transcribe speech, but also has other capabilities. As the focus of this thesis is on the speech recognition abilities of Whisper, the language identification and translation abilities will not be explored further. The capabilities of Whisper are:

- English speech recognition
- Multilingual speech recognition. The model is able to recognise 96 different languages. The performance of the language is based on the amount of training data available for that language.
- Translation tasks, the Whisper architecture is able to translate form English to a target language, and from a chosen language to English
- Language identification, the option to identify the language being spoken
- Voice activity detection

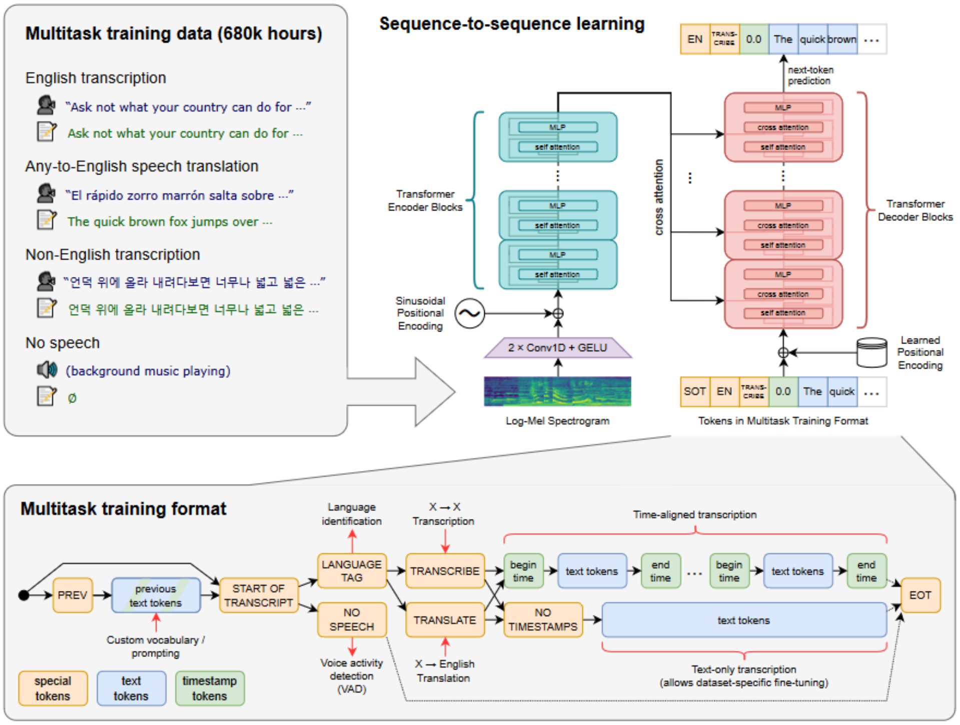
Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
⋯
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
⋯
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 2: The Whisper architecture (Radford et al., 2022)

Whisper is trained on transcribed audio that can be found on the internet. In total the Whisper architecture was trained on around 680 000 hours of labeled data, one of the largest datasets in supervised speech recognition (Radford et al., 2022). The Whisper ASR model is presented as a robust system. This means their model is able to generalise to many different datasets, without having to finetune the model first. Evaluating on a dataset the model was not trained on, is called zero-shot evaluation.

The authors didn't use the gold-standard for transcripts, which is human-validated transcripts (Radford et al., 2022). They chose this approach as creating a dataset of this size, using the gold-standard would be very time and labour intensive. Instead, Whisper uses weakly supervised learning. Transcripts made by ASR systems are detected and removed from the dataset. There are also multiple filters in place to filter out poor quality transcriptions and improve the overall quality of the dataset. Although these filters are used, the data used is likely lower quality then datasets in fully supervised learning (Radford et al., 2022).

There are different sizes of the Whisper model available. This was done to study the scaling properties of the model (Radford et al., 2022). The sizes available are Tiny, Base, Small, Medium, Large and Large V2. Some of these models are available in both an English and a multilingual version, the larger models only have a multilingual version. The Large V2 model was trained for 2.5 times more epochs compared to the large V1 model. The authors also added multiple methods for regularization. Regularization is used to prevent the model from overfitting, this is when a model when the model works well on training data, but does not properly work on the new test data. The following parts were added in the large V2 model: SpecAugment, Stochastic Depth and BPE dropout. The different models are summarised in table 1 (Radford et al., 2022). This table does not include the Large V2 model, as this model was released at a later time compared to the other models.

Table 1: The different Whisper models available (Radford et al., 2022)

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

For the multitask format, multiple different tokens are used. These tokens are used to specify the task at hand for the model (Radford et al., 2022). The tokens that can be predicted are tokens to start the prediction, the language being spoken, this token can be 'no speech' if there is no speech, a token to specify the task, whether to include timestamps and a token to indicate the end of transcript. These tokens are visible in figure 2 (Radford et al., 2022) as orange blocks.

In the paper different results for multiple different datasets are reported. These are zero-shot evaluations. On the LibriSpeech clean dataset (English) Whisper's large V2 model achieves a WER of 2.7%. This is comparable to the performance of the current baseline of supervised ASR models (Radford et al., 2022). The average WER on English datasets of the Large V2 model is a WER of 12.8%.

During this thesis the medium sized Whisper model will be used. This model was also tested on multiple multilingual datasets. Multiple of these multilingual datasets include Dutch data. The performance of Whisper on the Dutch part of multiple datasets is summarized in table 2 (Radford et al., 2022). The choice for the Whisper medium sized model was made based on computational limitations. The medium sized model is the largest model which would run on the available setup. Finetuning on larger models (the Large or Large-V2 model) would result in a 'CUDA out of memory' error.

Table 2: Reported performance of Whisper medium sized model on Dutch datasets (Radford et al., 2022)

Dataset	**WER**
Multilingual LibriSpeech	11.7%
Common Voice 9	8.0%
VoxPopuli	14.9%
Fleurs	9.9%

3.6 The encoder-decoder Transformer

An encoder-decoder Transformer was introduced by Vaswani et al. (2017). It consist out of multiple encoder and decoder blocks. Vaswani et al. (2017) and Radford et al. (2022) use different terminology for certain parts of the encoder and the decoder. For example, where Vaswani et al. (2017) uses the term *multihead attention*, Radford et al. (2022) uses multiple self-attention blocks. In the following explanation the terminology from Vaswani et al. (2017) is used. The encoder-decoder Transformer is visualised in figure 3 (Vaswani et al., 2017).

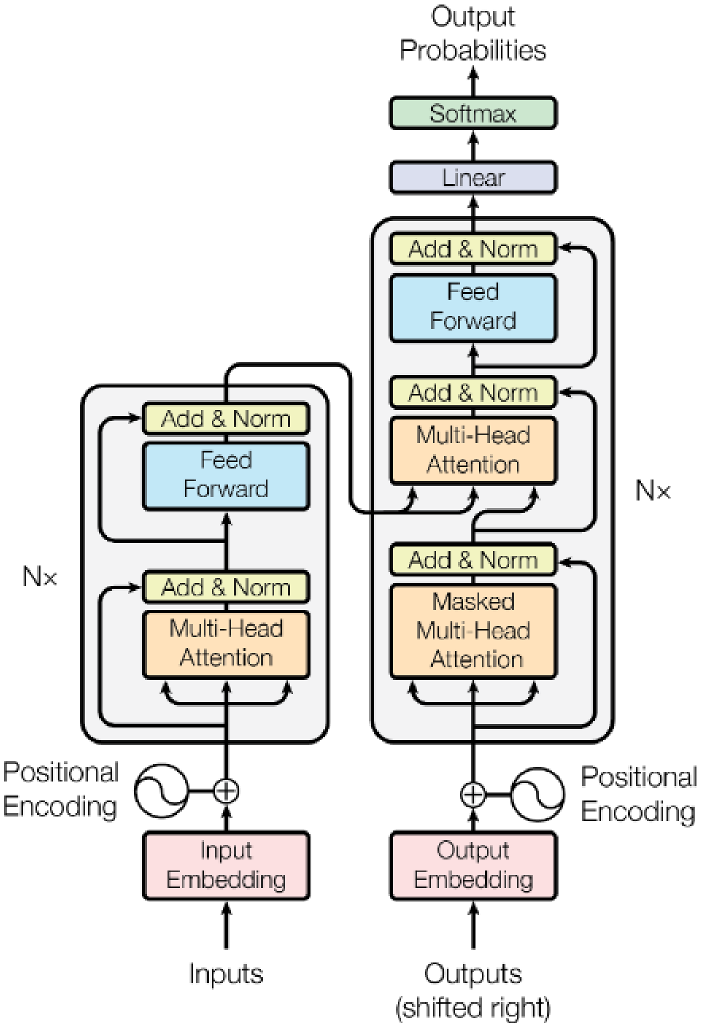

Figure 3: The encoder-decoder Transformer (Vaswani et al., 2017)

The input for the model is an audio sample. The audio signal is split in chunks of 30 second and is resampled to 16000 Hz. This is converted into a log-Magnitude Mel spectogram. In the Whisper architecture, the input representation is processed by the encoder with two convolution layers and a Gaussian Error Linear Units (GELU) activation function (Radford et al., 2022). This activation function was first proposed by Hendrycks and Gimpel (2016).

The spectogram is converted into *input embeddings*. The embeddings are a numerical representation of the audio [5]. This gives the model information on the word, the type of word and/or the meaning of a word. All features of the pre-training dataset are normalized, so all features are on the same scale, by scaling the input between -1 and 1, with an approximate mean of zero.

Next, the *sinusoidal Positional Encodings* are used to add positional encodings to give information on the sequence order/the position of tokens in a sequence. This is because the model itself does not have recurrence or convolution (Vaswani et al., 2017). These positional encodings are added to the input embeddings. The positional encodings together with the input embeddings, make up the input for the model.

The input from the sinusoidal positional encoding is in then passed to the encoder. The encoder consists of multiple encoder blocks. These blocks consist of self attention network and a Feed Forward Network. The *Self Attention* is used so the model can take into consideration, or pay attention to, the context of a sentence.

The output of the self attention adds a representation of the sequence by relating different positions of a sequence (Vaswani et al., 2017). Attention allows the model to focus on important parts of the sequence during the learning (Kamath et al., 2019). The model pays attention to the parts of the input, in the case of Whisper the audio, that are important. In the Transformer, the attention is calculated multiple times, performed in parallel. This is next concatenated into a final attention score (Vaswani et al., 2017). This is why it is called Multi-Head Attention.

In the *Feed Forward Network*, in the Whisper architecture called the MLP (Multilayer Perceptron), the information flows in one direction. In an MLP there is an input, an output and at least one hidden layer. The MLP is fully connected, the output of each layer is connected to the input of the next layer (Kamath et al., 2019). In the Transducer this Feed Foward Network is used to prepare the output of the attention layer for the next encoder block.

The output of the model is the transcribed sentence. During the training of the model, the target sentence is passed to the decoder. This sentence is changed into a numerical representation and given the positional encodings, this is the same process as the encoder input. The decoder also takes the output of the encoder as it's input. The decoder's self-attention works similar to the attention of the encoder, focusing on the output instead. The output is in this case the sentence. The cross-attention is where the attention is between the output of the encoder and the decoder, in the paper by Vaswani et al. (2017) this is called 'encoder-decoder attention'.

Because a lot of the computations are done at the same time, masking (the Masked Multihead Attention) and offsetting are used, this way the model can only use the available information at a given point in time (Kamath et al., 2019).

In the end, the decoder will output probabilities for word predictions. The highest probability score is the word that will be predicted.

[5] Getting started with Embeddings. https://huggingface.co/blog/getting-started-with-embeddings

3.7 Previous research on Dutch speech recognition

In this section multiple Dutch and Flemish datasets are named. The datasets CGN, COPAS and Domotica-3 will be used for this thesis and are discussed in depth in the methodology section.

3.7.1 Dutch speech recognition

Before we can look at Dutch dysarthric speech recognition, it is important to look at ASR performance for typical Dutch speech. Recent research has focused on both hybrid and E2E ASR systems. A couple of studies and their results are explained more in depth in the following section.

Van Dyck, BabaAli, and Van Compernolle (2021) trained and evaluated a hybrid ASR model on 155 hours of the Flemish part of the CGN (All components except for A, C and D) and the N-Best 2008 corpus using a Kaldi (Povey et al., 2011) implementation. This model is a HMM based hybrid system, using a TDNN acoustic model. The best WER achieved on the CGN development set is 12.4%. The best WER achieved on the N-Best 2008 corpus 10.12%. Exact details for the model and setup can be found in Van Dyck et al. (2021).

Röpke, Radulescu, Efthymiadis, and Nowé (2019) trained a DeepSpeech architecture on the entire CGN and a version using transfer learning from the English LibriSpeech dataset. DeepSpeech is an E2E speech system that uses a RNN at the core of its system (Hannun et al., 2014). The best processing technique found for the CGN was the removal of files containing overlapping transcriptions. The best WER achieved by Röpke et al. (2019), training fully on the CGN, is 30%. Using transfer learning between an English model helped to reduce this WER to 23.0%

Whisper medium sized model reports the performance on Dutch speech recognition in multiple multilingual datasets, these results are summarised in table 2 (Radford et al., 2022). When evaluated on the CGN, the Whisper large V2 model achieves a 12,27% WER. This is lower than the highest reported WER by Whispers large V2 model on the multilingual dataset, the VoxPopuli dataset with a WER of 12.9%. On the Whisper medium sized model a WER of 15.49% was found when evaluating on the CGN. This is slightly higher than the performance of the multilingual datasets found in table 2. When comparing this performance to previous research, the Whisper large V2 model has a similar performance to Van Dyck et al. (2021) on the CGN, although different components of the CGN were used in the evaluation. When compared to Röpke et al. (2019) the Whisper model outperforms the DeepSpeech architecture trained on the CGN and the model achieved by using transfer learning.

3.7.2 Dutch dysarthric speech recognition

There have also been multiple studies looking into the performance of ASR on Dutch dysarthric speech and how to enhance this performance. A couple of studies are explored slightly more in-depth.

Yılmaz et al. (2019) researched use of both articulatory features and acoustic features on Dutch dysarthric speech recognition. To achieves this, different acoustic models were compared. The ASR model used is a context dependent GMM-HMM system. One of the models used the Dutch and Flemish parts of the CGN (spontaneous conversations (component A), interviews (component F), discussions/debates (component G) and read speech (component O)) for training, and both the normal and the dysarthric speech of the COPAS sentence task for testing. The best WER achieved 29,0% on the dysarthric speech of the COPAS corpus. Exact details of the model can be found in Yılmaz et al. (2019).

Wang et al. (2021) explored an E2E Dutch dysarthric Spoken Language Understanding (SLU) system using Dutch normal and dysarthric speech data. A TDNN acoustic model is trained and the hidden layers are transferred as bottle neck features for the SLU system. The TDNN acoustic model is pretrained on the CGN (all components, except for the narrow-band recordings (component C and D), the spontaneous conversations (component A) and the simulated business negotiations (component E)). Next the model was finetuned using different speakers of the COPAS corpus, based on their intelligibility scores. The model is tested using the different speakers from the Domotica database. They found the WER for dysarthric speech for most speakers is above 35%, for one speaker within the high severity group the WER is as high as 80%. Finetuning on the full COPAS improves the recognition of speech with a lower intelligibility score. For speakers with moderate or mild dysarthria, the performance gain is smaller or the performance is even worse when compared to only pretraining on the CGN.

During the writing of this thesis, a study on Dutch dysarthric speech recognition comparing the Whisper model to different types of ASR architectures as the acoustic model to be used for an E2E SLU system was published (Wang et al., 2023). the small sized Whisper model was used. Although the aim of the research was to measure SLU performance, the WER for the different model types is reported. The results are summarised in table 3 (Wang et al., 2023). These models were pretrained on the CGN and finetuned on the dysarthric data from the COPAS dataset. The Domotica dataset was used for a zero-shot evaluation.

Table 3: WER per severity group for each model as reported in (Wang et al., 2023)

Model	**Severe**	**moderate**	**Mild**	**Mean**	**STD**
TDNN	51.54	31.26	39.40	40.73	10.21
Transformer	56.83	61.69	43.22	53.91	9,57
Whisper	53.81	40.40	37.51	43.91	8.70
XLSR-53	59.25	46.09	43.52	49.62	8.43

4 Research question and hypothesis

Current ASR systems have difficulties transcribing dysarthric speech data. This makes this technology less accessible for people suffering from dysarthria. The Whisper language processor is advertised as a model that is robust and usable out of the box on a lot of different datasets, this means there is no need to finetune the model. This makes it interesting to see if this is also applicable to dysarthric data. In this research the following research question is asked:

Will an end-to-end, weakly supervised Transformer based ASR model outperform an existing hybrid ASR model a recognition task on Dutch dysarthric speech data?

The Whisper architecture shows good results on zero-shot evaluations, also on the Dutch language (table 2). As the authors mention the models trained transfer well to existing datasets, the hypothesis is that the Whisper architecture will outperform the existing hyrbid ASR model.

When looking at the results reported by Wang et al. (2023), the Whisper small sized model performed comparable to a hybrid and other E2E models. It might be possible the medium sized model returns better results compared to the small model. Vielzeuf and Antipov (2021) and J. Li et al. (2022) report that E2E models are able to receive state of the art results. If the hybrid model and the Whisper model yield similar results without the need for finetuning, the Whisper model could show potential to be used for dysarthric speech without the need big datasets to finetune the model.

In the paper by Radford et al. (2022) mentions the need to study finetuning the authors expect that it is likely the performance can be improved by finetuning. In this study the results will show if finetuning with a small dataset of variable speech yields better results when compared to existing models. The hypothesis is that finetuning will increase the ASR performance

ASR models tend to perform worse on dysarthric data compared to typical speech. The amount of dysarthric speech available in databases is much smaller than normal speech (Yılmaz et al., 2019). The dataset used in the training of the Whisper architecture is constructed from transcribed audio found on the internet. predicted by the amount of training data for a language (Radford et al., 2022). The expectation is that there is less dysarthric data than typical speech on available the internet. This means Whisper will have less training data for dysarthric speech compared to typical speech. Previous research on Dutch dysarthric speech recognition showed a higher WER on dysarthric data then typical speech (Yılmaz et al., 2019; Wang et al., 2021). The expected result is that the performance on dysarthric speech will be worse when compared to typical speech.

5 Methodology

To answer the research question, the Whisper architecture will be compared to an existing hybrid implementation. The models will be evaluated on the Domotica-3 dataset (Ons, Gemmeke, et al., 2014). The results are reported based on severity of the dysarthria, which is based on the intelligibility scores: mild, moderate or high. These intelligibility scores are reported by the creators of the Domotica-3 dataset. With these results we can show if the weakly supervised end-to-end transducer model is able to outperform the hybrid ASR system in a Dutch dysarthric speech recognition task. The metric used for comparison is the WER calculated in percentages.

The following three models will be compared:

- The Whisper medium sized model without finetuning. This model is used to evaluate Whisper's out-of-the-box performance on dysarthric speech.
- The Whisper medium sized model, finetuned on all dysarthric data from the COPAS paired with transcripts
- A TDNN acoustic model, with a HMM-GMM model for audio feature alignment as reported by Wang et al. (2021)

The selection for the medium sized Whisper model was made based on computational limitations. The medium sized model is the maximum size model the available GPU is able to run without running into out-of-memory errors. The model consist of 4 layers, a width of 1024, 16 heads and 769M parameters. The full overview of the available Whisper models can be found in table 1.

The hybrid model used for comparison is the model proposed by Wang et al. (2021). In this research they built a TDNN acoustic model. For the audio feature alignment a HMM-GMM model trained in Kaldi was used (Povey et al., 2011). Wang et al. (2021) designed a Spoken Language Understanding (SLU) system for Dutch dysarthric speech. For this thesis's purpose, we will only look at the ASR system, and disregard the SLU part. This is because this part is not relevant for the research question.

For the pretraining of the model, Wang et al. (2021) used all Flemish components of the CGN, excluding component A, C and D. As component E does not include any Flemish data, it is assumed this component isn't used either. Next the model was finetuned on all dysarthric data from the COPAS. This dysarthric data was combined with 4,86 hours from CGN data.

In contrast to Wang et al. (2021), in this work the Whisper model will not be pretrained on Dutch speech using the CGN. This is because Whisper already performs well on the CGN dataset. To avoid overfitting on the CGN database, the pretraining step will be skipped.

5.1 Datasets

Multiple datasets are used during this research, these datasets are explained more in depth. For the selection of the datasets the choice was made to only select publicly available datasets. Other inclusion criteria were: the dataset needs to include transcriptions and needs to be in Dutch. For the finetuning and evaluation the dataset has to include dysarthric speaker data.

5.1.1 Corpus Gesproken Nederlands

The Corpus Gesproken Nederlands (CGN) (CGN, 2014) is a corpus of Dutch as spoken in the Netherlands and Flanders. The goal of the CGN was to develop a corpus for the Dutch language to be used for the development of speech and language technologies and contains around 900 hours of speech. The corpus contains fifteen different speech categories, ranging from read speech to spontaneous speech to interviews and debates (Van Eerten, 2007). The file structure of the CGN contains the audiofiles in the wav format, and the transcriptions in multiple formats. All the transcriptions are made by people. The different components of the CGN are summarised in table 4 (Van Eerten, 2007).

Table 4: Components of the Corpus Gesproken Nederlands (CGN, 2014)

Component	Type of data
Comp - A	Spontaneous conversations
Comp - B	Interview with teachers of Dutch
Comp - C	Spontaneous telephone dialogues (recorded via a switchboard)
Comp - D	Spontaneous telephone dialogues (recorded on MD with local interface)
Comp - E	Simulated business negotiations
Comp - F	Interviews/discussions/debates (broadcast)
Comp - G	(political) Discussions/debates/meetings (non-broadcast)
Comp - H	Lessons recorded in the classroom
Comp - I	Live commentaries (broadcast)
Comp - J	News reports
Comp - K	News (broadcast)
Comp - L	Commentaries/columns/reviews (broadcast)
Comp - M	Ceremonious speeches/sermons
Comp - N	Lectures/seminars
Comp - O	Read speech

5.1.2 Corpus Pathologische en Normale Spraak

The data set used for the finetuning is the Corpus Pathologische en Normale Spraak (COPAS) (COPAS, 2011). The COPAS is a Flemish corpus containing both typical and pathological speech. It was developed as part of a project to develop a speech technology based assessment tool, and to train speech language pathology students on pathological speech. The corpus contains 319 different speakers, 122 'normal' speakers and 197 pathological speakers. The pathological speakers have different underlying pathologies, 75 of these speakers have dysarthria. The corpus contains recordings of the Dutch Intelligibility Assessment, read speech, isolated sentences and spontaneous speech (Van Nuffelen, De Bodt, Middag, & Martens, 2009). The different pathologies of the speakers present in the COPAS is summarised in table 5 (Van Nuffelen et al., 2009). For the finetuning, the parts of the corpus containing dysarthric speakers and containing transcriptions was used. The following parts were used: The Dutch Intelligibility Assessment, sentence 1 & 2, text and text Marloes.

The Dutch Intelligibility Assessment (De Bodt et al., 2006) consist of three subtests containing 50 consonant-vowl-consonant words (Van Nuffelen et al., 2009). As mentioned before, this test is used in a speech therapy setting to determine an intelligibility score. Sentence 1 is the sentence 'Wil je liever de thee of de borrel?' (Do you prefer the tea or a (alcoholic) drink?). Sentence 2 is 'Na nieuwjaar was hij weeral hier' (After New Year's he was here again). The 'text' is one of eleven possible texts read out by different speakers. The text Marloes is a standardized text, often used in clinical practise when working with people with dysarthria (Van Nuffelen et al., 2009). Not every component was recorder for every speaker.

Table 5: Speakers pathology of the copas (Van Nuffelen et al., 2009).

Speaker Pathology	**Number of speakers**
Normal	122
Dysarthria	75
Hearing impairment	29
Laryngectomy	30
Cleft	38
Articulatory Disorders	17
Voice disorder	7
Glossectomy	1
Total:	319

5.1.3 Domotica

The Domotica corpus is a dataset containing Dutch dysarthric speech. The corpus contains sentences which can be used for home automation. There are four different Domotica corpera (1 - 4). The dataset that will be used for evaluation is the Domotica 3 corpus (Ons et al., 2014). Domotica 2, 3 and 4 are available via [4]. One example of a sentence of the Domotica 3 corpus is 'Aladin, deur dicht van badkamer' (Aladin, close bathroom door). The Corpus contains 17 different speakers, each with their own speakerID. These speaker ID's are consistent across the different Domotica datasets. Two speakers, speaker 31 and 37, are excluded in this thesis, as these speakers are children. This is in line with Wang et al. (2021) The speakers are divided by severity of the dysarthria based on their intelligibility score (IS). The IS was calculated using an automated procedure (Ons et al., 2014). The classification of the speakers can be found in table 6 (Wang et al., 2021). All sentences recorder by each speaker are used for evaluation.

Table 6: Speakers of the Domotica corpus (Wang et al., 2021).

Severity(IS)	**SpeakerID**
Mild (>85)	17, 40, 43, 44, 48
Moderate (70-85)	28, 29, 34, 35, 46, 47
High (60 - 70)	30, 32, 33, 41

[4] The Domotica Corpus. https://www.esat.kuleuven.be/psi/spraak/downloads/

5.2 Data preprocessing

In order to make the selected corpora suitable for the model, multiple different pre-processing steps were taken. The Whisper architecture accepts a maximum file length of 30 seconds. As most files in the CGN are longer than 30 seconds, the files are split up into chunks of this maximum size. Certain files of the CGN contain multiple speakers in the same file, the multiple speakers are split up, each resulting file containing the speech of one speaker. The COPAS and the Domotica-3 corpora contain audio files with one speaker at a time, this means the splitting of speakers wasn't necessary. If the files are longer than 30 seconds, these are split to fit the Whisper model's maximum file length. In the CGN, inaudible words are transcribed as 'xxx'. Files containing these inaudible words are omitted from the evaluation set. The Domotica dataset used 'ns' when there was no speech present, these transcripts were also removed.

Although the Whisper architecture has a normalization function, this isn't optimised for the Dutch language. For languages other then English, one general normalizer is used. The in the English normalization function, shortened forms are written out, certain spelling is corrected/changed and numbers are written in words instead of numbers. The general normalizer does not do this. The transcribers of the CGN used shortened forms in their transcriptions. Examples are m'n instead of mijn, z'n instead of zijn or 'ns instead of eens. Shortened forms are changed to their written out form as much as possible. It is however possible some shortend forms were missed. After the model predicted a transcription, the general normalizer is ran. This normalizer removes punctuation marks and capital letters from both the target reference transcription and the newly generated hypothesis. This way, the reference and hypothesis can be compared to one another.

During early evaluations of the Whisper models by the creators, the model would often output incorrectly guessed speaker names. Radford et al. (2022) predicted this is because the information on the spelling of the name is often not predictable from thirty seconds of audio. The Domotica-3 dataset contains the name 'Aladin' as the start of every transcript. In the first evaluations of the Domotica dataset, 'Aladin' was often misspelled in different ways. Because this doesn't necessarily reflects the ability of the Whisper model to transcribe dysarthric speech, the decision was made to correct common misspellings of the word 'Aladin' made by the system. The corrections made, are the same for each file. The spellings that are corrected: Aladdin, Alladin, Alla din, Allerdings, Allerding, Aladine, Aladina, Alla Dien, Allah die, Allah den, Alla die, Alla den and Alandin. Please note that this is not every misspelling of the name 'Aladin', as it would be impossible to go through every generated transcript to determine if this is a genuine misspelling or a wrong prediction. This means some misspellings may have been missed. Numbers are also corrected to their written out form, as the normalizer does not do this automatically for the Dutch language.

Another known problem in the Whisper architecture, is that the model can get stuck in a loop, transcribing the same word over and over, or having 'hallucinations' of completely unrelated transcriptions. According to the authors of the Whisper paper, these are due to a combination of failures in different parts of the decoding strategies (Radford et al., 2022).

An example of these hallucinations found while evaluating on the Domotica-3 dataset, would be 'Aadiedoe allee ta ta ta ta ta ta ta ta ta ta ta ta ta ta ta..' Where the 'ta' would be repeated over 220 times. Another example, would be where the model repeats the sentence 'Ik heb het niet goed.' (I got it wrong) 14 times, while the target sentence was 'Aladin, gordijnen neer in de living' (Aladin, close the curtains in the living room). And last, sometimes the transcripts would be completely unrelated to the target sentence. The model would output 'Wat? Wat hebben we een oordeel aan? Wacht, ik had het niet gezien. Ja, het is toch onbetaald...' ect. (What? What do we judge? Wait, I haven't seen it. Yes, it is unpaid...) continueing on for at least 75 words, while the target sentence was 'Aladin hoofdeinde op stand 1' (Aladin, head of the bed on position 1).

The choice was made to remove transcriptions containing these hallucinations. This choice was made because, when the model would transcribe the same word for 200+ times, this would have too big of an influence on the final WER measured. This is mainly because the datasets used for evaluation are quite small. For example, for one speaker, removing the hallucinations resultsed in a 100+ percentage point drop in WER.

5.3 Finetuning setup

The data used for the finetuning is taken from the COPAS corpus. The parts from speakers with dysarthria, containing transcriptions are used. This includes the data from the Dutch Inteligiblity Assesment, sentence 1, sentence 2, text and text Marloes. The dysarthric data was coupled with a small amount of data from the CGN. This data was taken from all components, excluding components A, C, D and E. This totals to a little over three hours of dysarthric data combined with 1,5 hours of typical Flemish speech used for finetuning. This gives a total finetuning dataset of around 4.5 hours.

The Whisper medium sized model is finetuned using Huggingface Transformers [5]. The medium sized Whisper model is finetuned on a single Nvidia A100 GPU for 750 steps with a learning rate of 0.00001. A batch size of 32 was used. Due to the small finetuning dataset, dropout was also implemented. The amount of steps was chosen based on the validation WER, after 750 steps this WER started to increase again. When trained for more steps, the model would output blank transcriptions. These blank transcriptions are discussed further in the 'Discussion'.

[5] Huggingface transformers: https://huggingface.co/docs/transformers/index

6 Results and discussion

6.1 Results

6.1.1 Corpus Gesproken Nederlands

The Whisper Large V2 model was evaluated on the CGN. This step was taken to evaluate whether pretraining on the CGN was necessary. The evaluation was done on all components of the CGN, except for component A, C, D and E, Whisper's large V2 model achieves a 12,37% WER. On only component O (read speech) Whispers V2 model achieves a 7,25% WER. On the medium sized model, a WER of 15.49% was found. The filtering of the hallucinations was not done in this step. This choice was made because this dataset is much larger, meaning the hallucinations would have less of an impact on the total WER. When scanning through the hypothesis and references, the following common errors were spotted. Please note that this is not a formal error analysis and these are not all the errors made, simply errors that are made often:

- *Word compositions* Some words that are usually written as one word in Dutch, are written as two or more different words in the hypotheses. Examples include (reference/hypotheses):
  - inwendig/in wendig
  - doorhad/door had
  - tweemaal/twee maal
  - niettemin/niet te min
- *Names* The read text contain a lot of different names. As names can be spelled in different ways, the hypotheses made by the model is often different than the spelling chosen by the transcribers of the CGN. The authors of the original paper mention this happening in their pre-training due to many transcripts in the pre-training containing the speaker names, and there is not enough context present in the audio files to predict the proper spelling (Radford et al., 2022)
- *Grammatical errors* A lot of errors in word conjugations are visible. Examples include (reference/hypotheses):
  - zuchtte/zuchten
  - stootten/stoten
  - predikten/predikte
- *Spelling errors* Multiple spelling errors can be noticed. One very noticeable mistake is the confusion between ij and ei. Sometimes resulting in non-existent words, or in a different wordtype. For some words, like hijgend/heigend, the model is very consistent in this error. Examples include (reference/hypotheses):
  - hijgend/heigend
  - nijdig/neidig
  - loutere/lautere
- *Not transcribing the full sentence* Sometimes the hypothesis is cut short. Not sure why this happens as there seems to be nothing wrong with the audio files. The files are of an appropriate length and are references are correct.
- *Style/shortened form difference* Where the transcribers of the CGN often chose the shortened form of certain words, the Whisper model does not.
- *Wrong transcription* The model simply picked a different/wrong word when transcribing.
- *Numbers* The hypotheses tend to write numbers using numbers, where the CGN transcriptions use words, e.g. one versus 1. This isn't corrected in the normalization function, as the build in normalization function only seems to be optimised for English.

6.1.2 Domotica-3 dataset

The results on the Domotica-3 dataset were collected before and after finetuning the model. The results before and after finetuning the Whisper medium sized model are summarized in a boxplot in figure 4. This boxplot compares the WER in percentages for the medium sized Whisper model with and without finetuning, categorized by severity. The Whisper medium sized model is present in the legend as 'no_finetune', the model after finetuning on the full copas is present in the legend as 'finetune_copas' The exact WER percentages are reported in table 7, 8 and 9. In these tables the comparison to the TDNN acoustic model proposed by Wang et al. (2021) is also made.

The results of the TDNN acoustic model reported by Wang et al. (2021) were made available via GitHub[6]. These are the results of finetuning on the full dysarthric COPAS data. The authors did not report on the results for each speaker, because of this, some results are excluded. For every speaker, ten different WER percentages are reported in the repository. For this comparison, the lowest WER for each speaker is reported.

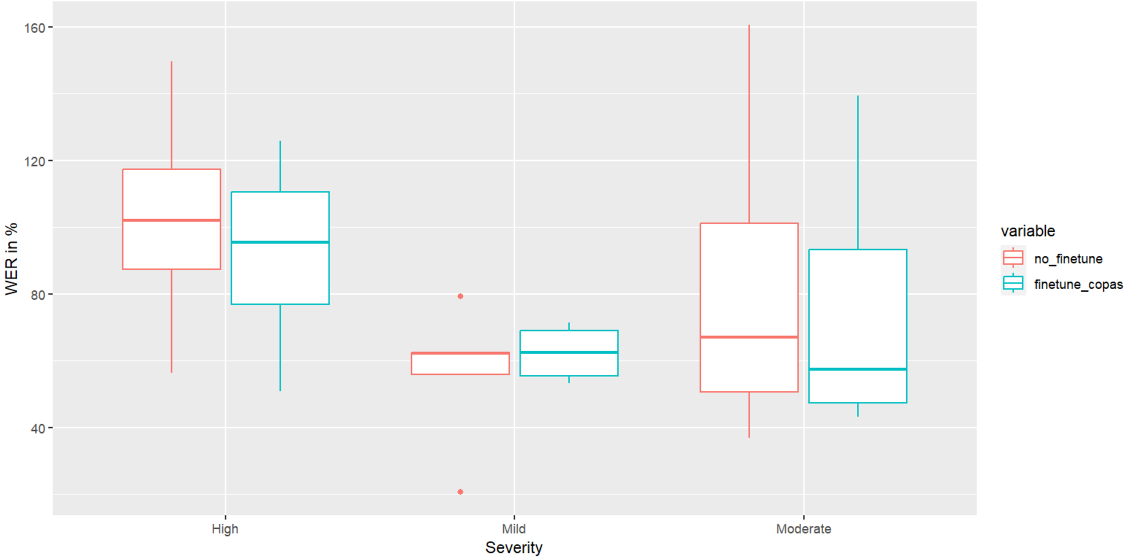

Figure 4: Results on the Domotica dataset

The ASR performance of the speech of fifteen different speakers was evaluated. For eleven speakers the WER was reduced after finetuning, while four speakers showed an increase in the WER. Five speakers showed a decrease in WER greater than 10 percentage point. The biggest increase in WER is speaker 44 (+32.62), and the biggest decrease in WER is speaker 33 (-23.97). Only two speakers achieved a WER below 50% after finetuning.

In speakers with a mild severity (IS > 85) dysarthria, the WER without finetuning varies from 20.60% to 79.24%. After finetuning this varies from 53.22% to 71.37%. Three out of five speakers showed a increase, and two speakers showed a decrease in the WER. The results are summarised in table 7. The mean increase in WER for the mild severity group is 6.39 percentage point. This means that overall, the ASR performance for the group of speakers with a mild severity dysarthria gets worse after finetuning the model.

For speakers with moderate severity (IS 70-85), before finetuning, the WER varies from 36.88% and 160.56%. After finetuning the WER varies from 43.14% to 139.44%. Only speaker 29 showed a increase in the WER. The other four speakers showed a decrease of the WER. The results are summarised in table 8. For moderate severity group, the mean decrease of the WER is 7,30 percentage point. Overall the ASR performance in the finetuned Whisper model for this group got better.

[6] https://github.com/wangpuup/Pre-training-with-dysarthric-speech

In the group of speakers with a high severity (IS 60-70) dysarthria, the WER without finetuning falls between 56.39% and 149.64%. With finetuning, the WER varies from 50.78% to 125.97%. The results are summarised in table 9. All speakers showed a decrease in het WER. The mean WER decrease is 10.70 percentage point. This again means the overall ASR performance for this group increased.

When comparing the results found by Wang et al. (2021) to the Whisper medium sized model and the finetuned Whisper model, the TDNN acoustic model outperforms both other models for every speaker. The biggest performance difference can be found in the moderate severity group. One example is speaker 34, where the difference in performance between the finetuned Whisper model and the TDNN acoustic model is 67,63 percentage points.

Table 7: Performance on mild severity, WER measured in percentages

Speaker	PP17	PP40	PP43	PP44	PP48
Whisper medium	55.87%	79.24%	62.48%	20.60%	62.22%
Finetuned on COPAS	71.47%	68.97%	55.36%	53.22%	62.35%
TDNN	37.09%	-	30.53%	-	26.48%

Table 8: Performance on moderate severity, WER measured in percentages

Speaker	PP28	PP29	PP34	PP35	PP46	PP47
Whisper medium	160.56%	36.88%	110.82%	71.93 %	46.92 %	62.14%
Finetuned on COPAS	139.44%	44.00%	104.97%	57.04%	43.14%	57.89%
TDNN	52.29%	-	37,34%	-	28,57%	-

Table 9: Performance on high severity, WER measured in percentages

Speaker	PP30	PP32	PP33	PP41
Whisper medium	106.50%	56.39%	149.64%	97.75 %
Finetuned on COPAS	105.35%	50.78%	125.97%	85.69%
TDNN	-	-	71,14%	61.73%

Next follows a short scan of the type of errors the model most commonly made before and after finetuning for the speaker with the biggest increase in performance (speaker 33), the speaker with the biggest decrease in performance (speaker 44). Please note that these results are from a short overview, and the exact amount of times an error was made, was not calculated. This is not a formal error analysis, as a full error analysis is outside the scope of this project.

For speaker 33 both before and after finetuning the WER was above 100%. This means the model still get nearly every word wrong or transcribes more substitutions when compared to the amount of deletions when compared the target sentence. Before the finetuning many sentences would start with ‘Ik ben’ or ‘Ik heb’ (I am/I have), after finetuning this changes to misspelling and misinterpretations of the name ‘Aladin’.

Examples, translations given below:

Target sentence -> before finetuning -> after finetuning:

‘Aladin slaapkamerdeur open’ -> ‘Deze is de slaapkammer door de open’ -> ‘Het slaapkamerdeur open’

(Aladin, open bedroom door’) -> (This is the bedroom through the open), containing spelling errors -> (The bedroom door open), using the wrong article.

‘Aladin hoofdeinde op stand twee -> ‘Ik heb geen idee wat hier nog te zien is’ -> ‘Aardien hoeft te heden op stand twee’

(Aladin, head end on position two) -> (I have no idea what’s left to do here) -> (Aardien has to present on position two), unlogical sentence.

In the transcriptions for speaker 44 different spelling errors can be seen similar to the errors found when evaluating on the CGN (e.g. hoofdijnde/hoofdeinde). Before finetuning the spelling of hoofdeinde was correct, after finetuning it was spelled hoofijnde. This might be a coincidence, but it does influence the WER. Other spelling erros can be found as well. There are also multiple blank transcriptions present in the output, which might influence the WER found. The finetuned Whisper model also seems to make more errors with regard to the name ‘Aladin’ when compared to the out-of-the-box Whisper model.

6.2 Discussion

When comparing the results found for the out-of-the-box, medium sized Whisper model, the Whisper model finetuned on the entire COPAS and the TDNN acoustic model, the Whisper architecture does not outperform the TDNN acoustic model for any of the speakers from the Domotica-3 dataset. There is a large difference in performance measured between the TDNN acoustic model when compared to both Whisper models, with difference between WER for certain speakers being as high as 67.63 percentage point in favor of the TDNN acoustic model. Although finetuning the Whisper medium sized model reduces the WER for eleven out of fifteen speakers. The increase in performance is limited for most speakers, with only five speakers having a more than 10 percentage point decrease in WER. In the group of speakers that showed a decrease in performance, two speakers showed a increase in WER of more then 10 percentage points, with the highest increase in WER being 32.62 percentage points when comparing the Whisper medium sized model to the finetuned Whisper model.

The results show that finetuning of the Whisper medium sized model on a small amount of dysarthric data does not seem to benefit every speaker with dysarthria. In this setup speakers with a mild dysarthria would even suffer from worse performance when the medium sized Whisper model is finetuned on the full COPAS. This is somewhat consistent with the findings of Wang et al. (2021), who saw limited improvements for speakers with an IS of 75 to 85, and an adverse effect for speakers with an IS of above 88. The speakers with an IS of above 88 are speaker 17 and 43. While speaker 17 indeed shows an increase in the WER, speaker 43 shows a decrease in WER when finetuning the Whisper model. After the finetuning, the WER of the different speakers seems to get closer together. The WER of the speakers performing best without finetuning increases, and the WER of the speakers with the highest WER decreases.

The results for the Whisper model and finetuned Whisper model show a big variability in the WER for speakers within the same severity group, with one speaker from the high severity group reaching a WER as high as 149.64%. As the WER is calculated by the number of substitution, deletions and insertions divided by the number of words in the target transcription, if the model predicts more words than there are words in the original transcriptions, the WER can turn out higher than 100%. This happens if the predicted transcript has more insertions than deletions.

If the speakers with a high WER speak in a slurred, slowed or segmented manner, the model might have difficulty segmenting the different words. This would be consistent with the findings by Moore et al. (2018), who mentioned that ASR systems can interpret slower or more segmented speech as the beginnings of new words. This could be one possible explanation for the high WER of a couple of speakers when compared to other speakers within the same severity group. As the speech of the different speakers of the Domotica-3 dataset was not analysed, it can't be said if this is indeed the case.

As mentioned, finetuning the Whisper medium sized model shows a large reduction of the WER for certain speakers. But with 13 out of 15 speakers having a WER of above 50%, these models are not suitable in their current form to be used in dysarthric speech recognition. Although performance gain found for most speakers in this thesis minimal, with some groups even showing a decrease in performance, the decrease of the WER of at least 10 percentage points for some speakers, using very limited data and finetuning for very few steps could show potential to be investigated further.

One important question to ask however, is if more dysarthric data might have improved the performance on dysarthric speech more. As this small dataset shows a degrading performance for some speakers with dysarthria, training on larger datasets might degrade this performance for these speakers even further and might make the model less accessible to non-dysarthric speakers.

When comparing the performance of the Whisper medium sized model without finetuning on the Domotica-3 dataset to the different multilingual datasets reported by Radford et al. (2022), see table 2, the performance for every speaker of the Domotica-3 dataset is worse compared to the multilingual datasets. All speakers also perform worse when compared to the performance on the CGN. This could be because of the dataset used, or because the Whisper model is unable to transfer all of its knowledge on the Dutch language to dysarthric speech, due to the difference between typical and dysarthric speech. Making the model fail to perform well on dysarthric data. One possibility is that, although the Whisper model is trained on a very large amount of data, this training data might include very little dysarthric data. As mentioned before dysarthria can lead to fear of communication and avoiding certain speaking contexts (Lowit & Kent, 2016). This combined with the fact that there are more people without dysarthria or other speaking disorders when compared to people with dysarthria, it could be possible there is very little transcribed dysarthric data available on the internet, this would mean the Whisper model has not been trained much dysarthric data. The authors of the Whisper paper mentioned that the performance of the Whisper model is tied to the amount of training data present (Radford et al., 2022). If the model is not able to properly transfer the knowledge of the Dutch language to the dysarthric speech due to the variability in dysarthric speech, combined with the possibility of little training data in the dataset, it could explain the poor results on the dysarthric datasets compared to typical speech.

The types of errors the Whisper model makes in Dutch speech recognition, mainly in the evaluation on the CGN, could also show that the language model for Dutch is probably not very strong. One example is the common error between 'ei' and 'ij' the model seems to make often. These combinations sound the same, but the spelling is dependant on the specific word. If multiple spellings are available, this is usually a different word type. The Whisper model is trained on audio paired with transcripts on the Internet. It can be expected the model did not learn these spellings directly from the training data, but likely learned the associated sound via the spectograms. The Whisper model does not seem to have a strong enough language model for the Dutch language to correct this or have learned the right spelling for the right word.

Another possible explanation for the limited performance gain on for some speakers with dysarthria, could be the limited amount of finetuning steps used, the size of the finetuning dataset and the variability in severity in the finetuning dataset. For the speaker included in the COPAS corpus, the IS varies from 28 to 100. For speakers in the moderate and high severity groups, this small amount of finetuning does improve performance, but does not translate well to the speakers within the mild severity group, who see a decrease in performance.

For the evaluation dataset in this research, the Domotica-3 dataset was used to evaluate the performance of the models. The disadvantage of this dataset is that every sentence contains the word 'Aladin'. This means that the models ability to transcribe this name correctly, or write it wrong semi consistently for filtering options, has a big impact on the performance. The Whisper model had difficulty with transcribing this name. As mentioned in the original paper, the spelling of a name is commonly not referable from 30 seconds of audio (Radford et al., 2022). This might influence the results of this thesis. Some of the spellings were corrected, but it was not feasible to go through every every transcript generated to determine if this was a misspelling of the name 'Aladin' or if the model had predicted something wrong. This would also would be a subjective decision which would make a fair comparison to the other models more difficult. As every utterance was quite short, with each audio file containing one sentence, this could have a big impact on the final WER. It is not known how the TDNN acoustic model handled the spelling of different names.

One problem during the finetuning was when it was noticed that the Whisper model would to output blank transcriptions when trained for too many steps or with a seemingly wrong setup. All data used to evaluate the model would result in an empty string. Unfortunately it wasn't possible to find more information on this phenomenon online or in different papers. One possible theory could be that this behavior might occur due to the inability of the model to generalise the very variable dysarthric speech, because of this the model reports every audio file as silence. It can not be said for certain if this is the reason for the blank transcriptions. Another possible explanation could be that the Whisper model usually processes chunks of audio of 30 seconds. Most of the files in the training are between 15 and 25 seconds long, but there are also files that are 2 to 5 seconds long. A theory would be these short files, train the model on too much silence. Evaluating on the short files from the Domotica-3 dataset does not seem to be the cause of any problems, as when evaluated on longer files, the model would also output blank transcriptions. The amount of blank transcripts produced was reduced by increasing the batch size, reducing the number of steps and adding a dropout function to the model. This does however mean that the finetuning was done for only 750 steps. As after these steps the evaluation WER started to increase and the amount of blank transcriptions produced also increased. 750 steps was found to be a good middle ground with optimal performance. In the first iteration, which trained for 4000 steps, the model would only produce blank transcriptions.

Although the amount of blank transcriptions were reduced in the final iteration of the finetuned Whisper model, the behavior of producing blank transcriptions still occurred on some of the files. The amount of time this happens is limited however. These blank transcripts might slightly influence the WER reported for the finetuned Whisper model. The amount of blank transcriptions produced was different between speakers, with some speakers showing no blank transcripts while others showed multiple. It is not known why this difference between speakers occurs as the file length and the sentences spoken are similar for each speaker. As this behavior only happened after finetuning the model, the choice was made not to remove these transcripts. Unlike the hallucinations, which happened in both the out-of-the-box Whisper medium sized model and the finetuned Whisper model. The hallucinations also had a bigger effect on the WER compared to the blank transcriptions. Including the hallucinations could increase the WER more than 100%, for example, speaker 33 would have a 277% on the Whisper model without finetuning if the hallucinations were not removed. In both the out-of-the-box model and the finetuned Whisper model, the transcripts containing hallucinations were removed. Although the exact numbers are not reported, it seemed these hallucinations occurred less after the finetuning step.

In short, the Whisper architecture is able to transfer some knowledge of dysarthric speech recognition to a new dataset after finetuning. The performance gain is however very limited and speaker dependant, with some speakers seeing big improvements while others speakers see a decrease in performance. For all speakers the performance is worse compared to the typical speech.

7 Conclusion and future research possibilities

7.1 Future scope

To make sure ASR systems are accessible to more people, more research needs to be conducted on dysarthric speech recognition. As a follow up to this thesis there are multiple directions future research could take with regards to Dutch dysarthric speech recognition, with or without the Whisper architecture. To determine different possibilities, a couple of suggestions based on this thesis are discussed below.

The Whisper architecture is build, in part, as a multilingual ASR model. In this thesis, the Dutch dysarthric speech recognition was explored, using only Dutch datasets. It would be interesting to explore the possibilities of combining dysarthric data in different languages to see if this enhances the ASR performance in these languages, using transfer learning. This would increase the amount of dysarthric training data available. This might also help to somewhat overcome the disadvantage in dysarthric speech recognition of limited availability of datasets, as the datasets available for dysarthric data are much smaller compared to databases for typical speech (Yılmaz et al., 2019). It should also be taken into account that this might not enhance performance for every speaker as seen in this thesis and in the study by Wang et al. (2021), where the performance of speakers with a mild severity dysarthria decreases due to finetuning on dysarthric speech. It might also be interesting to research the possibilities for personalised ASR using the Whisper architecture, if enough speaker data is available. Another possibility would be to train the model on speakers within the same severity group as the target speaker. This allows for more data to be used when compared to personalised ASR, where only the data of one speaker could be used.

Another possible direction for future research is to increase the performance on Dutch speech recognition, before finetuning on Dutch dysarthric data. The performance on the CGN by the Whisper model was determined to be good enough to not need any pretraining on the CGN, but it might be beneficial to pretrain on typical Dutch speech before finetuning on dysarthric speech to see if this decreases the amount of spelling and grammatical errors made by the Whisper model. In the results found by Wang et al. (2023), a study published during the writing of this thesis, the Whisper architecture showed better results compared to the results found in this thesis. The methodology by Wang et al. (2023), mainly pretraining on the CGN, might be better suited for the Whisper model and be able to achieve better results.

In the results there was a short look at the type of errors made by the Whisper architecture in both the recognition of typical speech and dysarthric speech. A full error analysis was not within the scope of this thesis. One possibility for future research could be to do a full error analysis on dysarthric speech recognition to see what the model is not able to transcribe correctly. The Whisper model would perform very different on speakers within the same severity group, and the severity scores of the Domotica-3 dataset were calculated using an automated process (Ons et al., 2014). To further evaluate why ASR systems struggle to transcribe dysarthric data it would be very interesting to evaluate the speech of dysarthric speakers on multiple ASR systems and have multiple different speech language therapists evaluate on severity and characteristics of these speakers. By combining the performance rates with the different characteristics of the speech and the type of erros made by the systems, it might be possible to better report on which characteristics make dysarthric speech recognition difficult for ASR systems. As of right now, mainly the abnormality of speed or the severity of the dysarthria are mentioned as causes of the degradation in performance. Knowing more on this might enable creators of ASR systems to research ways to enhance the performance for these speakers with dysarthria.

The authors of the Whisper paper suggested to improve the decoding strategies of the Whisper model to reduce the amount of hallucinations or loops the model tends to get stuck on (Radford et al., 2022). These hallucinations and loops had a big impact when evaluating on small datasets. Reducing these hallucinations and loops could increase the performance of the model.

And last, due to computational limitations, the choice was made to use the medium sized Whisper model for this experiment. There are more pretrained checkpoints available of the Whisper architecture. The results found in this thesis might improve by using the Large or Large-V2 model released by the creators of the Whisper model.

7.2 Conclusion

Some people might struggle to communicate clearly. For some people this is because they suffer from dysarthria. Current ASR systems perform worse on dysarthric speech when compared to typical speech. Multiple reasons for this worse performance can be found in the literature. Including less available data for dysarthric speech, the large variability in dysarthric speech, and the abnormality of speed present in dysarthric speech. Good ASR performance is important to make technologies that use ASR systems more accessible for people with dysarthria. To contribute to the research being done in this field, in this thesis, the following research question was asked:

Will an end-to-end, weakly supervised Transformer based ASR model outperform an existing hybrid ASR model a recognition task on Dutch dysarthric speech data?

The hypothesis presented was that the weakly supervised Transformer based ASR model would be able to outperform the hybrid ASR model due to E2E models being able to receive state of the art results (Vielzeuf & Antipov, 2021; J. Li et al., 2022). And the Whisper small sized model performing comparable to other hybrid E2E models (Wang et al., 2023) when evaluated in a Dutch dysarthric speech recognition task. The other hypothesis was that the model would perform worse on dysarthric data when compared to typical speech. This is because generally ASR systems tend to show this difference in performance.

The answer to this research question asked is, that with the setup used, the end-to-end, weakly supervised Transformer based ASR does not outperform the hybrid ASR model. For the weakly supervised Transformer based model, the Whisper architecture (Radford et al., 2022) was used. The medium sized checkpoint was used and finetuned on the COPAS and evaluated on the fifteen speakers of the Domotica-3 dataset. Both datasets contain audio from Flemish dysarthric speakers. These results were compared to the TDNN acoustic model, with a HMM-GMM model for audio feature alignment as reported by Wang et al. (2021)

The finetuned Whisper model did not outperform the TDNN acoustic model on any of the speakers present in the Domotica-3 corpus. For speakers with a mild severity dysarthria, the ASR performance decreased after finetuning. For the speakers with moderate of high severity dysarthria, a limited increase in performance was seen. Possible explanations for this limited performance could be: the finetuning configuration used, the blank transcriptions produced by the model, the dataset used for evaluation or the size of the dataset used for finetuning.

These results show that the first hypothesis, that the weakly supervised Transformer based ASR model would outperform a hybrid ASR system, is rejected. This is because the weakly supervised Transformer model could not outperform the hybrid ASR model on this specific Dutch dysarthric speech recognition task. The hybrid model outperformed the Whisper model on every speaker, before and after the finetuning of the model. The second hypothesis, that the model would perform worse on dysarthric speech compared to typical speech, is confirmed. When comparing the results of the Domotica-3 dataset to the results reported in table 2, no speaker of the Domotica-3 dataset achieves similar results to the multilingual datasets containing typical speech.

Although the Whisper model could not outperform the hybrid system, with the small amount of finetuning steps used during this thesis, some speakers still achieved quite good performance increase. With three speakers having a 10+ percentage point increase with a very limited amount of data used in the finetuning dataset. Even if this thesis shows a worse performance of the Whisper model, it is not certain if this is because the Whisper model performs worse on dysarthric speech, or if the used setup and methodology are not suitable for the Whisper model. The setup might not be ideal, due to the spelling and grammatical errors the Whisper model makes on typical Dutch speech, it would be recommended to pretrain a future dutch dysarthric speech recognition system on typical Dutch first to enhance the knowledge of the Dutch language. Another recommendation would be to use a different dataset used for evaluation. As the Domotica-3 dataset relies heavily on names in its target sentences. Another possibility might be also be that the Whisper architecture needs more data during finetuning to prevent the amount of blank transcriptions produced. This could be tested by adding more CGN data to the finetuning dataset.

The authors of the Whisper paper report that the Whisper model generalises well across different domains, tasks and languages (Radford et al., 2022). This claim does not seem to hold up to the Domotica-3 dataset specifically. As the out-of-the-box Whisper medium sized model, showed a poor performance on this dataset. More research is needed to see if this is also the case for different dysarthric datasets. To fully determine if Whisper architecture is usable for dysarthric speech recognition, more research is needed. This research can focus on finetuning the Whisper model on a larger dataset, personalised ASR, finetuning on speakers within the same severity category or a combination of these factors. As mentioned before, it seemed the Domotica-3 dataset was not the ideal dataset to be used for evaluating the performance of the Whisper model, due to the big reliance on names in the transcripts. In a different, future research it would be interesting to see if evaluating on a different dataset, containing less names, might yield better results.

It is important to continue the research for dysarthric speech recognition. This way this technology could become more acessible to this demographic, making sure everyone can use the new advantages of ASR. With future improvements to dysarthric speech recognition, possibilities will open up for using ASR systems in eldercare, healthcare or at home increase and allow the optimistic ideas and initiatives mentioned in the introduction to come to fruition.

References

Corpus Gesproken Nederlands - CGN (Version 2.0.3). (2014). [Data set]. (Available at the Dutch Language Institute: `http://hdl.handle.net/10032/tm-a2-k6`)

Corpus Pathologische en Normale Spraak (COPAS) (Version 1.0.1). (2011). [Data set]. (Available at the Dutch Language Institute: `http://hdl.handle.net/10032/tm-a2-n3`)

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, *12*(2), 246–269.

De Bodt, M., Guns, C., & Van Nuffelen, G. (2006). *Nsvo: Handleiding.* Vlaamse Vereniging voor Logopedie: Herentals.

De Russis, L., & Corno, F. (2019). On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, *5*(3), 163–172.

Duffy, J. R. (2019). *Motor speech disorders e-book: Substrates, differential diagnosis, and management* (fourth ed.). Elsevier Health Sciences.

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... others (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., ... Gruenstein, A. (2019). Streaming end-to-end speech recognition for mobile devices. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings v2019-May (2019 05 01)* (pp. 6381–6385). doi: 10.1109/ICASSP.2019.8682336

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jaddoh, A., Loizides, F., & Rana, O. (2022). Interaction between people with dysarthria and speech recognition systems: A review. *Assistive Technology*, 1–9.

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep learning for nlp and speech recognition* (Vol. 84). Springer.

Li, B., Chang, S.-y., Sainath, T. N., Pang, R., He, Y., Strohman, T., & Wu, Y. (2020). Towards fast and accurate streaming end-to-end asr. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6069–6073).

Li, J., et al. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, *11*(1).

Lowit, A., & Kent, R. (2016). Management of dysarthria. In *Aphasia and related neurogenic communication disorders* (pp. 527–556).

Moore, M., Venkateswara, H., & Panchanathan, S. (2018). Whistle-blowing asrs: Evaluating the need for more inclusive speech recognition systems. *Interspeech 2018*.

Mustafa, M. B., Rosdi, F., Salim, S. S., & Mughal, M. U. (2015). Exploring the influence of general and specific factors on the recognition accuracy of an asr system for dysarthric speaker. *Expert Systems with Applications*, *42*(8), 3924–3932.

Ons, B., Gemmeke, J. F., et al. (2014). The self-taught vocal interface. *EURASIP Journal on Audio, Speech, and Music Processing*, *2014*(1), 1–16.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.*

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Rao, K., Sak, H., & Prabhavalkar, R. (2017). Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 ieee automatic speech recognition and understanding workshop (asru)* (pp. 193–199).

Röpke, W., Radulescu, R., Efthymiadis, K., & Nowé, A. (2019). Training a speech-to-text model for dutch on the corpus gesproken nederlands. In *Bnaic/benelearn.*

Sluijmmers, J., Singer, I., Versteegde, L., Zoutenbier, I., & Gerrits, E. (2016). *Prevalentie en incidentie van dysartrie en spraakapraxie bij volwassenen.* Rapport voor NVLF van Lectoraat Logopedie Hogeschool Utrecht.

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

Van Eerten, L. (2007). Over het corpus gesproken nederlands. *Nederlandse Taalkunde*, *12*(3), 194–215.

Van Nuffelen, G., De Bodt, M., Middag, C., & Martens, J.-P. (2009). Dutch corpus of pathological and normal speech (copas). *Antwerp University Hospital and Ghent University, Tech. Rep.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vielzeuf, V., & Antipov, G. (2021). Are e2e asr models ready for an industrial usage? *arXiv preprint arXiv:2112.12572*.

Wang, P., BabaAli, B., et al. (2021). A Study into Pre-training Strategies for Spoken Language Understanding on Dysarthric Speech. *arXiv preprint arXiv:2106.08313*.

Wang, P., et al. (2023). Benefits of pre-trained mono-and cross-lingual speech representations for spoken language understanding of dutch dysarthric speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*(1), 1–25.

Yılmaz, E., Mitra, V., Sivaraman, G., & Franco, H. (2019). Articulatory and bottleneck features for speaker-independent asr of dysarthric speech. *Computer Speech & Language*, *58*, 319–334.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, *22*(2), 99–112.

A Appendix

The code used in the writing of this thesis was made available via GitHub [6]. This repository includes:

- The preprocessing code for the CGN, COPAS and Domotica-3 corpora.
- The code used to finetune the Whisper medium sized model
- The finetuned Whisper medium sized model
- The code used to evaluate the performance of the medium sized model

For further information, please refer to the GitHub repository.

[6] https://github.com/LianRUG/ThesisWhisper