university of groningen

campus fryslân

Fine-tuning ASR to specific noise environments: noise robustness in a climbing gym

Elja Leijenhorst

University of Groningen

Fine-tuning ASR to specific noise environments: noise robustness in a climbing gym

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. V. Verkhodanova (Campus Fryslân, University of Groningen)
and
Dr. M. Coler (Campus Fryslân, University of Groningen)

Elja Leijenhorst (s4979427)

July 31, 2023

Acknowledgments

This thesis has not been a smooth journey for me, and at times I have been close to giving up. I am proud to have completed this project, during which I have learned a lot and which I eventually also enjoyed a lot. I am incredibly grateful for having some amazing people by my side who convinced me to keep going. Most importantly being my mom, who has truly been my rock throughout this process by sharing many hours of studying and helping me to keep in mind what is important. Just as grateful am I for having the luck to be supervised by Vass, for her never-ending positivity, problem-solving, and the hot chocolates. If you would not have kept on believing in me, I honestly would not have made it to this finish line. I hope many more students may be inspired and cheered up by you.

I am also grateful for the kind help of Shekhar Nayak, who often gave creative input opening the door for new possibilities.

I feel lucky for the good contact with the staff of the climbing gym who provided me with the opportunity to use their facilities, together with all the climbers and employees with all their positive reactions and willingness to collaborate.

Lastly I would like to mention Qiushi Zhu, for their patience and kind emails, and providing me with the dataset of Prasad et al. (2021).

Contents

Glossary

Term	Definition
Autobelay	A belaying device which is situated on top of a climbing wall, which requires no human actions for safe belaying. A climber attaches the line to their harness using the metal locking system or a carabiner. When a climber lets go off the wall, the device slowly lowers them with a rattling sound.
Belaying	Belaying is the act of securing a climber with a rope and a manual piece of equipment, ensuring that they are lowered down safely when they fall or need to descend from the climbing wall. For lead climbing, catching a fall as a belayer can include jumping against the wall.
Bouldering	A climbing discipline where a thick foam landing mat is used for safety instead of ropes, with climbing walls up to a maximum of five meters tall. Climbers often jump down from halfway up the wall.
Clip 'n climb	A playground area consisting of plastic walls with autobelays on top, with rubber mats below. Mostly used by children, which goes hand in hand with lots of enthusiastic vocal expressions.
Lead climbing	A climbing discipline where the rope is not already attached to the wall, and the climber needs to push the rope through the quickdraws (making a sharp clipping sound). This is their last point of protection, causing them to land against the wall if they fall when climbing above the last attached quickdraw. This might result in a loud bang of feet against the wall, both from the climber and belayer's side.
Toproping	A climbing discipline where the rope is already attached to the top of the wall. The only action that the belayer needs to take is making sure that the rope stays tight, while the climber only needs to climb. When a climber falls, this is often not very audible, and the rope stretches dynamically to catch the fall.
Quickdraw	A piece of climbing equipment consisting of a metal carabiner connected to the wall by a strong strand of fabric. While climbing, the used quickdraws may move and clatter against the wall.

Abstract

This research aims to improve the noise robustness of automatic speech recognition (ASR), specifically in the context of climbing gyms. There is no known research on ASR performance in sports facilities, while these have been reported to often have poor acoustics (Wróbel & Pietrusiak, 2021). Sport climbing requires a safety course with personal guidance, which causes this sport to be poorly accessible for members of the deaf community. ASR could potentially be of help in these situations if it performs well enough in this loud environment. The goal of this thesis is to optimize an ASR model for the one specific acoustic environment of the climbing gym and explore whether this has an advantage over a general noise-robust ASR model. This study encourages the use of ASR in sports facilities and contributes to ASR noise robustness in general.

Following methods similar to Zhu et al. (2022) and Schlotterbeck et al. (2022), two wav2vec 2.0 models (pre-trained on an English LibriSpeech dataset of 960 hours) were fine-tuned on two LibriSpeech datasets mixed with different types of noise. One model is fine-tuned on speech mixed with newly created noisy background recordings from the climbing gym, while the other is fine-tuned on speech mixed with publicly available noises from daily real-world environments like restaurants and public transit stations (multi-condition training).

The noise-robust model fine-tuned on gym noise speech did outperform the general noise-robust model for speech from the climbing gym by a relative 6%. Noise robustness of both models improved with 48% in terms of WER, compared to the baseline model, demonstrating the effectiveness of fine-tuning on noisy data. These results suggest that fine-tuning an ASR system to a specific noise environment would have an advantage over a general noise-robust ASR system, posing an optional solution to a well-performing ASR application in the climbing gym.

Index Terms – Automatic speech recognition, wav2vec 2.0, fine-tuning, noise robustness, multi-condition training, sports facilities

1 Introduction

Interest in research and development in the field of automatic speech recognition (ASR) has been growing quickly in the last few decades and great progress has been made. Nowadays, ASR systems have achieved remarkably high performance of transcribing speech when it comes to well-represented languages and perfect acoustic conditions. However, in the real world, these perfect conditions hardly ever exist. Challenging factors such as different accents, different voices, background noises and other unexpected variables make it difficult for ASR systems to perform well on real-world data (Shrawankar & Thakare, 2013). Even though these issues are well-known, the vast majority of ASR systems is still trained on 'perfect' datasets. Many of these datasets consist of standard-accent read speech with minimal background noise, which requires the lowest processing effort for the model and do not realistically prepare the model for the real world (Szymański et al., 2020). Fortunately, research is branching out and attention towards these challenging factors is growing, and researchers are collectively working towards more inclusive ASR systems. This entails accounting for varying speaker characteristics, as well as accounting for variations in acoustic environments (Y. Wang & Gales, 2012).

In this thesis, I focus on the influence of noise on ASR performance. What possibilities are there to make an ASR model as noise-robust as possible? And is it always necessary for a model to be robust against all noises, or is a specific subset sometimes enough? An ASR model that is used only in a particular setting needs to perform well exclusively in that specific environment. For example, sports facilities are noisy settings, which have been found to often be a loud environment with a fairly constant set of noises and bad acoustics, due to the often high ceilings and hard smooth surfaces, which cause long reverberation times (Wróbel & Pietrusiak, 2021). In this research I zoom in on a specific climbing gym which fits this description (Mountain Network Noardwand in Leeuwarden, The Netherlands), to optimize an ASR model for this gym. The loud noise environment might prohibit some people from easy communication, while also hindering people who rely upon ASR for daily tasks. Furthermore, before being able to climb independently, people have to follow a safety course with important instructions and guidance. This course cannot be offered properly to deaf people by instructors who do not know sign language, unless an interpreter is present. ASR could potentially be of help in these and similar situations by providing live subtitles. Because sport climbing is quickly gaining popularity (NKBV (Royal Dutch Climbing and Mountaineering Federation), n.d.), the demand for accessibility and inclusivity in the community grows. This thesis contributes to encouragement of using ASR in sports facilities for a more inclusive sports environment, as well as to representation of sport climbing in literature. To my knowledge, this thesis is the first work targeting ASR in a climbing gym environment.

Research on noise-robust speech recognition has been carried out for other noise contexts and with multiple approaches (further discussed in Chapter 2). In Schlotterbeck et al. (2022) the fine-tuning of an ASR on noise from a Spanish classroom has been investigated, but there was no comparison between performance of models with different training setups. In this thesis, I build upon their research with similar methods in different noise environments and investigate the value of these methods. The purpose of this thesis is to explore whether fine-tuning an end-to-end ASR model with environmental noise recordings from a specific environment leads to improved ASR performance for that specific noise environment[1], or that another noise-robust model would perform equally well.

[1] To distinguish the different types of noise and avoid confusion, from this point on the term *noise* is used when referring to a specific sound produced by a distinct physical sound source, while the term *noise environment* is used when referring to the mixture of all sounds typical for an environment, consisting of multiple separate noises and reverberations.

1.1 Thesis Outline

To best answer this question, Chapter 2 elaborates on the case study and provides the necessary background information, including a review of earlier studies related to the topic. Subsequently the exact research questions of this thesis and my hypotheses following from the literature review are defined in Chapter 3. Chapter 4 then describes the methodology used to answer the research questions, followed by the experiment results and a discussion of these in Chapter 5 and Chapter 6 respectively. Lastly, I conclude by summarizing the takeaways from this study in Chapter 7.

2 Background and Related Works

This chapter contains background information on the topics introduced in the introduction and further explains how this thesis came to be. The chapter consists of five sections, which together lead to the specific research question and hypotheses which are presented in Chapter 3. In the first section of this chapter, I introduce the noise setting of the climbing gym together with the sport of climbing. Following this, I give a brief overview of ASR performance and solutions for noisy speech data. The chapter concludes with a section on wav2vec 2.0, the ASR framework that was used in this study.

2.1 Noise in sports facilities and a case study of the climbing gym

Sports facilities often have poor acoustics, due to the combination of high ceilings and hard surfaces, which cause long reverberation times. Wróbel and Pietrusiak (2021) examined the noise levels in commercial training facilities and gyms[2], and investigated ways to reduce the problem. Together with a large variety of sounds consisting of noise caused by sports activities and other background noise, bad room acoustics result in a loud noise environment. While Wróbel and Pietrusiak (2021) focused on the problematic situation for humans, who can experience the noise in sports facilities as disturbing or even suffer from hearing loss, this thesis primarily focuses on the influence that such noise has on ASR. Just as background noise can severely interfere with communication for people with hearing loss or people communicating in their second language, performance of ASR systems can drop when the SNR (Signal-to-noise ratio) is too low (Zhu et al., 2022). In this section, the noise environment of the climbing gym is presented, together with a brief recommendation for reducing noise in similar facilities. The relevant climbing-related terms are defined in the Glossary.

> **A brief introduction to sport climbing -** Sport climbing is a dynamic and social sport which is usually performed in pairs. Though its origin lies in rock climbing outside in nature, I concentrate on the indoor variant of the sport. A climber tries to reach the top of the wall using strength and technique, while their safety is being secured with a rope from the ground by a belayer. Both parties are attached to the rope: the climber with a knot on his harness, the belayer through use of manual belaying device. When the climber is high up in the air, they communicate loudly using so-called rope commands: a fixed set of short and clear statements that can be recognized easily even when there are many factors hindering communication. In the gym where collection of audio data for this thesis takes place, three types of climbing are practiced: toproping, lead climbing, and climbing with automatic belay devices. More information about these terms can be found in the glossary at the beginning of this thesis. There are significant differences in the three varieties in terms of the sound they cause, making the difference relevant to this research.

The specific noise environment of a climbing gym is made up of many different sounds, with a varying composition throughout the day and week. Most climbing gyms are commercial organizations open to the public. It is not unusual for gyms to host groups of beginners of all ages, and provide clinics, courses and training sessions. As the precise range of possibilities differs per climbing gym, this section is limited to the most usual situations found in a standard climbing gym. In the collection of

[2] Though I acknowledge that their research focuses on a different type of gym facilities than is treated in this thesis, their findings are valuable for this study on climbing gyms too. It should be noted that one cannot make a direct comparison and knowledge transfer, still many of the acoustic features described also apply to the majority of climbing gyms.

common sounds, there is a broad variety in terms of frequency (low-pitched and high-pitched sounds), periodicity and amplitude. Below, I give an (non-exhaustive) overview of the collection of sounds that make up the noise environment of the climbing gym:

- Clipping quickdraws
- Impact drivers being used to screw holds into the wall, at specific routesetting hours
- (Metal) materials clattering against each other
- (Metal) materials on wooden shelves or on the floor
- Footsteps and other regular human movement (rustling of clothing)
- Rope movements (occasionally a rope falling to the ground from ten to twenty meters of height)
- Bangs against hollow climbing walls (from climbers moving and falling against the wall)
- Continuous baseline background sounds such as: wind, echoes, ventilation system, lamps, heating
- Autobelays (automatic belay devices)
- Speech: regular conversation and shouting
- Music
- Reverberation of all sounds

In the data collection section of Chapter 4 I go into more detail about how these sounds accompany each other in the climbing gym of this study, and how it was ensured that all aspects of this noise environment were captured.

As mentioned at the beginning of this section, this facility and many other climbing gyms (as well as other sports facilities) have a naturally ‘loud’ design, with the hard surfaces and high ceilings. Next to improving our ASR systems, it would be favourable for both the technology and all of the people using the space, to simply reduce the noise of the environment. As many of the factors cannot be directly manipulated, it would be beneficial to improve the acoustics of the space, so that single noises have a smaller impact (Wróbel & Pietrusiak, 2021). This can be done using relatively simple techniques such as ensuring that the floors are carpeted or have rubber tiles (instead of concrete), placing coverage on walls that are not used for climbing, or placing structures with a high absorption coefficient (designed specifically for improving acoustics) just below the ceiling. The last technique has already been applied to many bouldering gyms in the Netherlands, which are often situated in old warehouse-like buildings primarily made of concrete and metal. Implementation of these techniques reduces reverberation time significantly.

Sport climbing is rapidly increasing in popularity around the world. In the Netherlands, the sport has been going through an explosive growth over the last twenty years (NKBV (Royal Dutch Climbing and Mountaineering Federation), n.d.). Where in the early nineties the number of facilities in the country could be counted on one hand, nowadays there are roughly eighty facilities for sport climbing and bouldering. Many new gyms have opened their doors and existing facilities are becoming increasingly crowded. Climbing is a high-risk sport which requires proper education on safety before one can independently engage in the sport. This education is provided in the form of a belaying course taught by a certified instructor. Belaying is a simple but delicate action which requires individual guidance and feedback before people can become proficient in this task. This course can often not be offered properly to deaf people, since sign languages are not commonly mastered by the speaking community. Because of this, sport climbing is generally not inclusive to the deaf community (Taub, 2022). ASR could potentially be of help in these and similar situations, by e.g. providing live transcriptions. Furthermore, people who need ASR to perform daily tasks should be able to continue doing so in a loud environment. This can range from, for example, people with reduced motor functions who benefit from using voice commands to control their electronic devices, to people with hearing loss who can benefit from reading transcriptions of what people are saying out loud.

2.2 ASR performance on clean and noisy recordings

Though speech recognition still strongly favors a specific set of perfect conditions like a native standard English accent, relatively slow and clearly pronounced speech, little background noise, and use of a good microphone, it has already come a long way in becoming more robust against changes in these conditions. A lot of research is being done on how to improve model performance by testing out different types of models and structures, and using different datasets for training. An important challenge that needs to be addressed in order to make ASR universally applicable, is to overcome the presence of noise, which in most real-world situations is an issue to some degree

Before zooming into the effect of the noise environment from the climbing gym discussed in Section 2.1 on ASR, I discuss the bigger picture through a summary of the general influence of varying conditions on performance of a 'normal' ASR model, i.e. an ASR model that has not specifically been optimized for noise robustness. Speech datasets are usually divided into different splits for training and testing (and validation), which are most often highly similar, causing ASR performance results to not be truly representative when a model is trained and evaluated on splits from the same dataset. Whereas humans can effectively understand speech in unseen noise conditions without any specific training, ASR systems on the other hand perform poorly when tested on such new noise conditions not seen during training (Lippmann, 1997). It is known that performance of machine learning models degrades when testing data strongly deviates from data seen during training. Unfortunately, this is common practice. Szymański et al. (2020) expressed criticism on this topic and studied the trustworthiness of word error rates (WERs) reported in research in the past five years. They tested a wide range of ASR systems on different datasets than those used for training and found that the average WERs where significantly higher than the WERs reported in the papers presenting the models. They outlined several problems with the methods of acquiring performance results, one of these problems being the usage of a split of the same corpus for testing as was used for training.

Likhomanenko et al. (2020) also broke this skewed pattern of using very similar train and test sets, by training and testing a number of ASR models on different datasets in order to investigate transfer across datasets and conditions. As can be expected, they found that 'normal' models scored significantly lower for noisy datasets compared to clean datasets. While Likhomanenko et al. (2020) mostly focused on domain transfer of non-noise-robust models for clean datasets, this might also generalize to transfer from different noise-robust models to different noise-robust datasets. In the following section, different approaches of making ASR robust to noise are being explored, elaborating on the method of choice used in this thesis.

2.3 Improving noise robustness

Multiple approaches have already been found to improving noise robustness of an ASR system. These techniques can be applied at three distinct levels of the ASR-system: noise-reduction or speech enhancement before training, robust feature extraction, or adapting the model (Van Segbroeck & Narayanan, 2013). Because the noise in the climbing gym is a complex composition of signals spread out over a wide range of frequencies and partly overlapping with speech frequencies, it is difficult to apply general speech denoising methods like those described by

Van Segbroeck and Narayanan (2013), where the contrast of periodicity between speech and noise is exploited.

Other methods that are focused on training, are pre-training on noisy speech (Likhomanenko et al., 2020), fine-tuning on noisy speech (Schlotterbeck et al., 2022, Likhomanenko et al., 2020, Zhu et al., 2022, (Prasad et al., 2021)), fine-tuning on noise-clean paired data (Maas et al., 2012) and applying attention (Higuchi et al., 2021).

One thing that can greatly influence the complexity and size of an ASR model is the range of domains in which it performs adequately. An ASR model that is always be used in a specific setting or for a specific goal, only needs to perform well in that specific setting or for that specific goal. To illustrate this, consider the following example of two ASR systems with very different goals. An ASR system for a cash dispenser suffices with the basic commands for the keys, and there is no need for it to make use of state-of-the-art technology and have the capacity to understand whole conversations. Additionally, since it is situated in a public space on the street, it should be noise-robust enough to filter out at least the usual traffic noises for that specific site like car sounds and voices. A smartphone voice assistant on the other hand, is used in more diverse settings and the user expects to hold a relatively natural conversation with their device. As the prompts are extremely diverse, the vocabulary should be very broad.

The case study of an ASR system in the climbing gym seems to have similarities to both of these examples. Regarding the noise robustness, it is always surrounded by a relatively constant set of noises. The model should be robust against quite a lot of noise, but only in that specific place. Therefore, there is no need for the system to have the capacity to filter out a huge variety of noises that are not likely to occur in the gym and it suffices to use local noise for training. As mentioned above, a general rule about artificial intelligence (AI) is: when training data is more similar to testing data, performance rises. For models that perform in a limited domain, as is the case in the current research, one can take advantage of this principle in order to optimize the model for its use. However, it should not be overlooked that if the training data is too specific or limited, the model is prone to being overtrained and to perform badly for new cases. Because the noisy data in this research is a complex composition of sounds with little regularity, overtraining is less likely to occur quickly. In fact, D. Wang et al. (2021) used the method of introducing random noise in training data to reduce overfitting and found it to improve performance compared to the baseline model and to other methods of overfitting like drop out, layer normalization and spectrum data augmentation.

2.4 Fine-tuning on noisy speech

A straightforward method for making an ASR model noise-robust is fine-tuning. This technique is essentially an extra training phase of an AI model. Instead of starting training with randomly initialized weights, the weights of a pre-trained model are taken as a starting point. Because the model does already have a fair representation of the target data from pre-training, a much smaller dataset of the fine-tuning domain is required for good results. This makes it suitable for the purpose of keeping the process of optimization as accessible as possible.

One study where fine-tuning is used as a method to improve noise robustness of an ASR model, is Schlotterbeck et al. (2022). They fine-tuned a large version of wav2vec 2.0 (pre-trained on the Spanish common-voice dataset) over a small six-hour dataset of 4th to 8th grade Chilean lessons in multiple school subjects. Their training and testing data was split into three datasets of the different school subjects. All three were used for training, and test results were reported separately for all three datasets for the fine-tuned model, the base wav2vec 2.0 model and two other popular cloud-based systems. Compared to the baseline models they found large performance improvements, relatively reducing the WER with 35%. With this result, Schlotterbeck et al. (2022) showed that transfer learning with limited annotated data can be used to improve model performance for that specific domain. They compared model performance on multiple varieties of noisy data, and at the same time compared multiple models. However, they did not compare performance between different models fine-tuned on the different varieties of noisy data, but only compared their fine-tuned model to non-noise-robust models.

A second study on improving noise robustness of ASR by means of fine-tuning is Zhu et al. (2022). They also performed an experiment where different ASR models were tested on multiple noise types. They found that pre-training wav2vec2.0 on noisy data leads to improved performance on noisy test sets, but performance for clean speech drops. By using a baseline model pre-trained only on clean speech and fine-tuning with noisy speech, the resulting model maintains a high accuracy when tested on clean speech. Whereas Schlotterbeck et al. (2022) recorded and transcribed speech in the classroom as training and testing data, Zhu et al. (2022) used an existing clean speech dataset and mixed this with noise recordings (speech from LibriSpeech (Panayotov et al., 2015) and noise from FreeSound (Font et al., 2013)). This method might cause data to be somewhat less realistic, but does obtain a similar effect of adapting the model to different circumstances and making it robust to noise. Furthermore, data preparation is far less time-consuming when noise and speech are mixed together if possible, avoiding the lengthy transcription process.

2.5 Wav2vec 2.0

Both Schlotterbeck et al. (2022) and Zhu et al. (2022) used a pre-trained version of wav2vec 2.0 as a baseline model. Wav2vec 2.0 (Baevski et al., 2020) is an open-source end-to-end model for automatic speech recognition. It is relatively easy to train using Hugging Face Transformers (Wolf et al., 2019), and many different versions of the model have been published, pre-trained on many different corpora for many different languages and types of speech[3]. An exceptional achievement of wav2vec 2.0 is that its large pre-trained version reaches a WER as low as 4.8% on the clean test set of LibriSpeech with only 10 minutes of labelled training data. This means that it is ideal for circumstances with limited data.

Wav2vec 2.0 is an end-to-end model that uses self-supervised learning for pre-training, taking unlabelled audio data in the shape of raw waveforms as input. After this, the supervised process of fine-tuning with labelled audio data is necessary for the model to learn representation between all the audio data and text. For this, wav2vec2.0 uses the connectionist temporal classification (CTC) loss function. After fine-tuning, the fully trained model can be used to predict transcriptions of new audio input similar to the data that has been used for fine-tuning. Because of this split training process, the model can be used even when little data is available. Wav2vec 2.0 can, for example, be pre-trained on a large dataset of multiple languages and fine-tuned on a small dataset of a low-resource language like Swahili (not included in training), resulting in satisfactory performance rates for Swahili, which could otherwise only be acquired with much larger datasets. The same holds for fine-tuning the model on data with other variables like non-typical voices or noise. When trained only on clean speech, model performance on noisy testing data is significantly reduced compared to when fine-tuned also on noisy speech (Zhu et al., 2022).

[3] See `https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec#wav2vec-20` and `https://huggingface.co/models?sort=trending&search=wav2vec2` for a broad collection of pre-trained and fine-tuned wav2vec 2.0 models.

3 Research Question and Hypotheses

The main research question central to this thesis is as follows:

> "Does an end-to-end ASR model fine-tuned on noise recordings from a specific environment outperform a general noise-robust model for that specific environment?
> In particular, will this be the case for the specific noise environment of a climbing gym?"

For this question, I hypothesize the following based on earlier findings from Schlotterbeck et al. (2022) and Zhu et al. (2022):

Primary hypothesis:

H1. I expect that the environment-specific ASR model will outperform the general noise-robust ASR model, when tested on data from the specific noise environment.

Secondary hypotheses:

H2. The general noise-robust model will outperform the environment-specific model on other noises.

H3. Both noise-robust models will outperform the base (non-noise-robust) model for both noisy test sets.

H4. The base model will outperform both noise-robust models for clean test sets, but not with a large difference.

If the hypotheses are not supported by the results of the experiments and both noise-robust models perform equally for both test sets, it is possible that the environmental noise is too diverse for the model to learn a good representation of the acoustic properties of the noise.

4 Methods

As explained in the background chapter, fine-tuning is the method that I used to answer the research question. To compare whether a model fine-tuned on a specific noise environment has an advantage over another noise-robust model, I created two datasets from different noise environments and fine-tuned the base model into two separate models with these datasets, following a methodology similar to the one used by Schlotterbeck et al. (2022). In this chapter, I explain in detail how exactly the noise datasets were created and discuss the experiment structure in more detail.

4.1 Data

One of the goals of this thesis is to keep the process of optimizing a model for a specific sound setting uncomplicated, to make it as accessible as possible. This means that the costs and resources required should be minimalistic without compromising performance. This way, the process could be easily applied to multiple locations. In a perfect scenario, owners of noisy locations like a climbing gym can optimize their own model by uploading their noise recordings to a user interface which processes it and returns the optimized model in a ready-to-use manner. Keeping in mind the goal of making the process low-effort, the process of recording audio data should also be simplified as much as possible. Recording training speech directly in the noisy environment might give the best model performance, but it is an elaborate and time-consuming process. An easier way would be to record the noise environment, and later mix that with an existing available speech dataset. Hence, this is the method that I used in this project.
The first dataset is created using environmental audio recordings from the climbing gym (hereafter referred to as *Gym Noise*), whereas the second dataset is created using a varied collection of noise recordings publicly available at FreeSound.org (hereafter referred to as *Public Noise*).

Figure 1: An overview of the climbing gym (Mountain Network Noardwand, Leeuwarden).

4.1.1 Gym Noise + LibriSpeech

The first and most important dataset that was created is built up from a collection of noisy audio recordings from the climbing gym, mixed with speech from the widely known LibriSpeech corpus (Panayotov et al., 2015). This is a speech corpus consisting of audio and transcriptions collected from

English audiobooks. The corpus is divided in training, testing and development sets, and recordings are categorized as ’clean’ or ’other’ based on recording quality. The model used in this thesis is pre-trained on the training set (train-clean-960), while I used the development set (dev-clean) for fine-tuning and the test set (test-clean) for evaluation. In this section, the full recording process of the Gym Noise is explained, along with relevant considerations that were made. I will also explain how the noise was added to the LibriSpeech datasets.

4.1.1.1 Collecting audio

Recording equipment For obtaining the Gym Noise, the TASCAM DR-100MKII Linear PCM Recorder was used. Of the microphones available for use at the University of Groningen, this was the most suitable for the purpose of the study. The DR-100MKII is a small, portable recorder with a wide variety of settings and the ability to obtain high-quality recordings. It has omnidirectional microphones, making it suitable for capturing sounds from all parts of the gym and being less dependent on the exact placement than a unidirectional microphone would be.

The recorder was set to the following specifications: WAV-format, 44.1kHz sampling rate, 24-bit precision. Even quiet sounds at the other side of the climbing gym were captured as realistically as possible with the 24-bit range, without losing quality by compressing the signal. The recordings were made in mono, to suit the requirements of the ASR model used for the experiments.

Before starting the data collection phase, a pilot recording was created to assess if the microphone was suitable, and to determine optimal microphone placement.

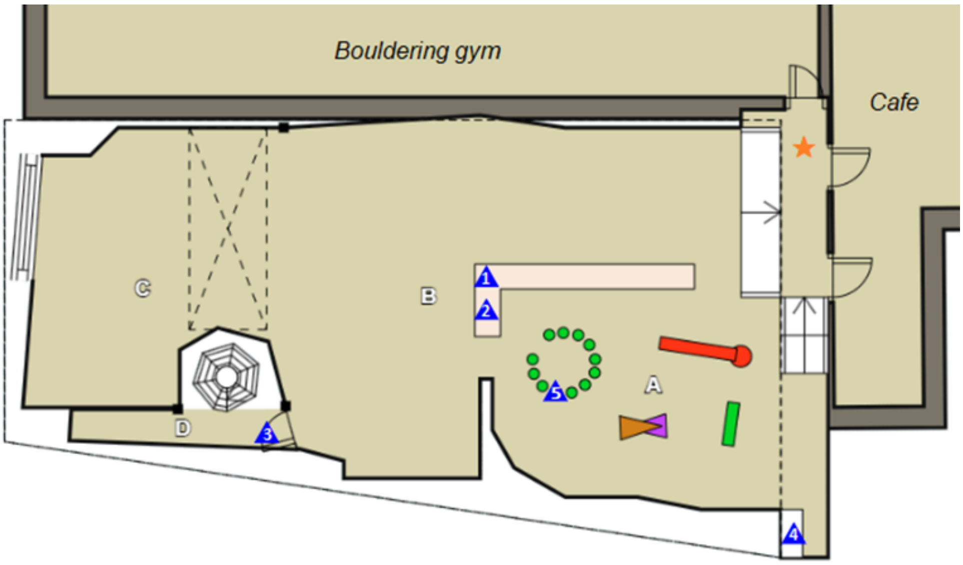


Figure 2: A floor plan of the gym, with the recording locations (1-5) and different areas (A-D).
A: Clip ‘n climb children’s area
B: Toproping area
C: Lead climbing area
D: Routesetting storage area
★: Photo location of Figure 1
▲: Microphone placements

Recording schedule The location of recording was the climbing gym Mountain Network Noardwand (Leeuwarden, The Netherlands) (Figure 1). In the process of creating a data collection plan, there were two important factors to keep in mind. One being the time of recording, given that the noise levels in the gym differ greatly throughout the week. The gym is usually opened from 10:00 a.m. to 11:00 p.m., which leads to different audiences using the gym at different times. Weekends and afternoons are filled with excited screams from children and usage of the auto-belays in the Clip ‘n climb area, while evenings only contain sounds from the actual climbing sport and mostly adult’s voices. Next to these general patterns there are also many irregular events taking place in the gym, ranging from busy school classes of teens to the weekly process of setting of new routes, which entails that climbing holds are being screwed into the walls using impact drivers. To get an optimal representation of all possible sound events, audio was recorded on a great variety of times and dates, planning in such a way that all aspects would be covered.

Microphone placement Next to this, considerations were made about the placement of the microphone. Figure 2 shows a floor plan of the gym. The optimal position would be in the centre of the gym, where all noises from all areas are optimally captured. However, since the gym can get rather busy, this is not always a safe place for the recording equipment. To be able to capture the average noise situation well and to include the different sound representations at different places, while also keeping the equipment safe, I was forced to place the microphone at a variety of places throughout the recording process. These are indicated with blue triangles on the floor plan in Figure 2.

On one day of recording, all areas were unexpectedly busy, leaving location 4 to be the only option left to safely place the microphone. However, when reviewing the recordings this turned out to be too far in the corner to properly catch any of the climbing sounds and led to a minimal representation at a low volume. Ultimately, the audio of this day was kept and used, because a certain portion of the noise environment was captured, and even though it was quite deviant from most of the data, it still contributes to the goal of creating a complete representation of the noise environment of the entire gym. Figure 3 depicts a distribution of recording hours per microphone location.

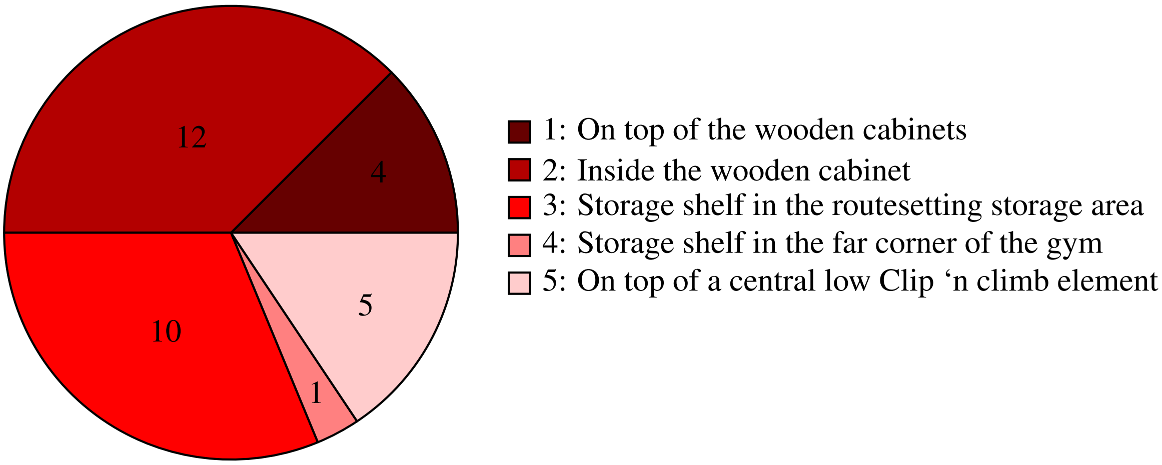


Figure 3: Distribution of recording hours per location of the microphone, corresponding to the blue triangles in 2

.

4.1.1.2 Inspecting and processing collected audio After a total of approximately 32 hours of data collection, a preliminary inspection of the Gym Noise was performed. From the sounds common to climbing gyms listed in section 2.1, some sounds especially stood out in the recordings. I observed the spectrograms of a part of the recordings in Praat (Boersma & Weenink, 2021) and identified several intensity peaks. The largest peaks were mostly low-pitched, often being bangs against the wall and the beat of music. The presence of music is a variable that could potentially be challenging for model performance, since every song has quite different characteristics, and music partly is in the same frequency range as speech. These factors might make it difficult for the model to distinguish all features from speech and music at times. Smaller peaks in the spectrogram were mostly high-pitched sounds from the clipping of carabiners.

As the climbing gym in this research offers climbing lessons for schools and kids' parties, some recordings contain a large amount of enthusiastic screams and children's voices. Random samples were observed to check for artifacts, and some loud intelligible conversations were cut out. This is clarified in the section on ethical considerations. After I reviewed these samples, an exhaustive corpus was formed, which consists of recordings that cover different parts of the day in an equal manner, as visualized in Figure 4.

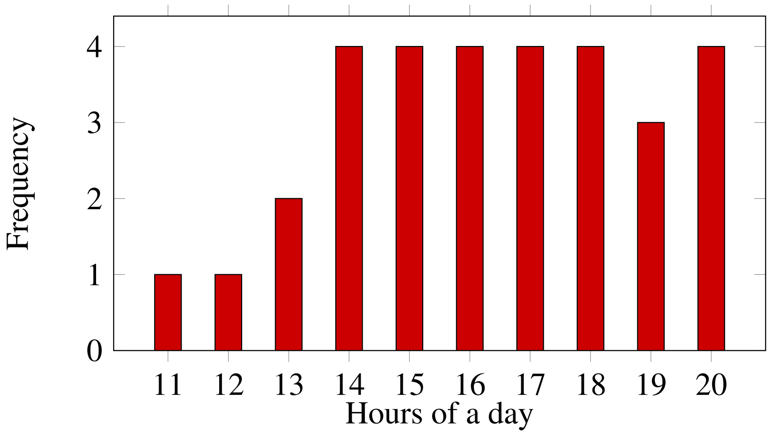


Figure 4: Total number of recordings per hour of the day

4.1.1.3 Mixing noise audio with clean speech For an optimal comparison with earlier literature (Zhu et al., 2022), the LibriSpeech dev-clean subset (5.4 hours) (Panayotov et al., 2015) and a selection of the test-clean subset were used for fine-tuning and evaluation respectively. Following the methods of Prasad et al. (2021) and Zhu et al. (2022), 120 files were randomly chosen from the test-clean subset for validation. These sets were then mixed with the Gym Noise. In this process, all audio was adapted to satisfy the requirements for the wav2vec 2.0 model from Hugging Face (converted into WAV format, converted from stereo to mono and downsampled to 16 kHz where necessary).

For each LibriSpeech file, a random interval with the same length was selected from the Gym Noise recordings, making sure that the noise fragment did not overlap with a previously selected interval. The speech file and the noise interval were then mixed without changing the amplitudes, causing the SNRs in the final dataset to be quite diverse. This choice was made to let the dataset resemble real-world situations as closely as possible; people speak at different volumes and different distances to the microphone, and background noise volume varies greatly per moment[4].

To complete the dataset, a metadata file was created containing the transcriptions of all noisy speech files, conforming the format suitable for Hugging Face datasets (Lhoest et al., 2021).

[4] The python script used for mixing and all other scripts used in this project can be found at `https://github.com/cmfuego/thesis_noiserobustASR`

4.1.1.4 Ethical considerations and privacy When capturing a good representation of all the common noises in the climbing gym, the presence of intelligible speech is inevitable. Since recorded speech can be used to identify a person, it is classified as personal data by the GDPR (General Data Protection Regulation). This means that the data must be managed carefully, and everyone involved in the recording process should be informed well and have control over their own personal data. To safeguard privacy, multiple measures were taken in accordance with the Ethics Committee of the University of Groningen (faculty Campus Fryslân). Firstly, some measures were taken to inform visitors of the gym of the situation:

- Climbers were informed of the recording situation by placing multiple information signs: posters in Dutch and English were put up at the wall where announcements are usually placed, at the check-in point, in the middle of the gym. A smaller sign was positioned close to the microphone. The information sign can be found in the Appendix. All employees were informed about the situation so that they could answer questions. By informing people beforehand, they were given the choice to avoid standing in a close proximity to the microphone.
- To reach as many people as possible, an informing message was posted in the Leeuwarden climbing community WhatsApp group (including most regular customers of the gym, with >250 members), along with a picture of the information sign and microphone, so that climbers knew what to look out for.
- Most likely, these methods do not allow to reach every person before their visit, and the presence of the microphone can easily be forgotten when people continue their usual routines. Therefore, every sign included my contact detail, and a QR-code leading to a form, where people could request to delete a certain time span of the recording or get in touch for other questions.

There were no negative reactions or requests to delete data. The only questions and responses were expressed in person in the gym, to employees and myself, which were all positive. Lastly, to minimize the amount of speech remaining in the final corpus and to ensure ethical processing of the audio, the following measures were taken. These were also mentioned on the information pamphlet.

- Microphone placement was chosen in such a way that most of the speech becomes background noise. Speech from people standing by the climbing wall fades into the background, making it largely unintelligible and causing the speaker to be (almost) unidentifiable.
- All parts of the recording containing obviously intelligible speech should be deleted or made unintelligible.
- All recordings will only be used for purposes of this thesis by the researchers, who only listen to a minimal amount of the recorded audio.

4.1.2 Public Noise + LibriSpeech

The second dataset is based on the dataset from Prasad et al. (2021), in order to follow the methods of Zhu et al. (2022) as closely as possible. This was again done using the LibriSpeech dev-clean subset and a selection of the test-clean subset (Panayotov et al., 2015) for fine-tuning and evaluation respectively. Prasad et al. (2021) selected a number of noise recordings from FreeSound.org, forming a varied set containing noises from public scenes like traffic and restaurants, continuous machine noise and babble noise. Training or fine-tuning on data from different noisy environments is known as ‘multi-condition training’ (Du et al., 2014).

For every speech file, a noise file was randomly picked, and the two files were combined with a randomly picked SNR from 0 to 20. Since there were less noise files than speech files, it was accounted for that each noise file at each SNR would be used before using a noise file at the same SNR for a second time.

Even though Prasad et al. (2021) included a ready-made test set (LibriSpeech test-clean files mixed with noise), I used this only as an example and changed the structure. Whereas in their dataset, each speech file was repeated to be combined with each noise at each SNR, I chose to make sure that no speech file was repeated for optimal comparison with the Gym Noise dataset. I did not change the selection of noises.

4.2 Experiment structure

Figure 5 shows a schematic representation of the experiment structure. The base model is fine-tuned separately on the two datasets, in the same configuration. As the baseline model for this thesis I used wav2vec2-base-960h[5] (Baevski et al., 2020), which is a wav2vec 2.0 model pre-trained and fine-tuned on the English LibriSpeech training data (960 hours of unlabelled audio). For fine-tuning, I followed Hugging Face's ASR guide[6] and trained the models for 59 epochs with a batch size of 8 and a learning rate of 1e-5[7]. Checkpoints were saved every 1000 steps, in order to choose the optimal number of steps.

After the fine-tuning process, the performance for all three models (base model, model fine-tuned on Gym Noise, and model fine-tuned on Public Noise) was tested on both noisy test datasets and a selection of clean speech (similar to the noisy test datasets, a random selection of 120 speech files from the LibriSpeech test-clean set).

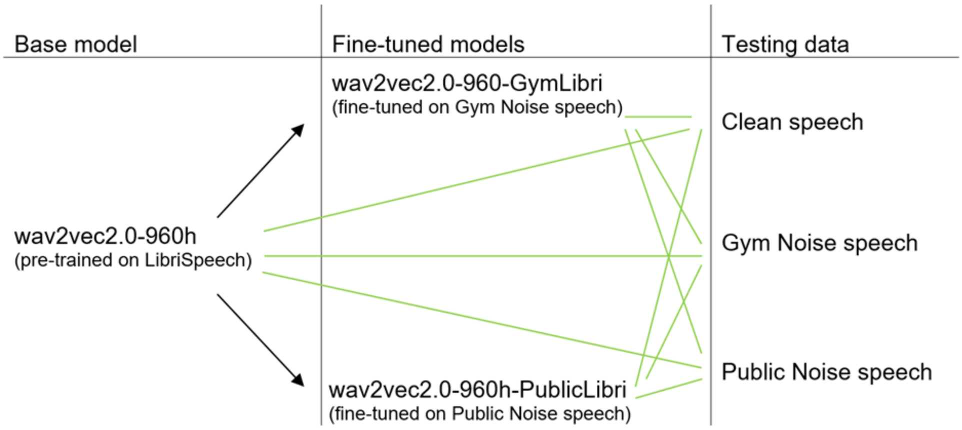


Figure 5: A schematic representation of the experiment structure.

4.3 Metrics: WER and CER

For a simple comparison with other literature, I use the word error rate (WER) and character error rate (CER) metrics to measure model performance. Especially the WER is a widely common performance metric in the field of speech recognition. It is used to compare the transcriptions of the model with the original validated transcriptions and calculate how large the difference between these two is. To

[5] The pre-trained model can be found at https://huggingface.co/facebook/wav2vec2-base-960h

[6] Hugging Face "Automatic speech recognition" how-to guide for fine-tuning and evaluation can be found at https://huggingface.co/docs/transformers/v4.27.2/en/tasks/asr

[7] The complete fine-tuning script can be found at https://github.com/cmfuego/thesis_noiserobustASR

be precise, the WER is calculated by adding all substitutions, insertions and deletions of words in the sequence together, and dividing the total by the total number of words in the original sequence (see Formula 1). The CER is closely related to the WER, with the only difference being that for CER, differences are calculated on character level instead.

$$WER = \frac{S + I + D}{N} \tag{1}$$

It should be noted that these metrics are not optimal for testing how much of the meaning is conveyed, as certain words are more important than others (Y. Y. Wang et al., 2003). To illustrate this, it might be the case that common words (such as articles or conjunctions) are recognized perfectly, while some important keywords might be transcribed incorrectly. The transcription might have a low WER, but still be unintelligible because the core of the text is missing. The other way around, when only the core words are transcribed correctly, the gist of the message can be understood even with a high WER. In other words, the diversity and ingenuity of language is not easily captured in a single number. I selected these as the metric of choice for best comparison with earlier literature, but it remains important to keep in mind the diverse performance factors that are not represented by this one simple number, such as which noises have the most impact on performance.

5 Results

After fine-tuning the models, I used the code included at the GitHub page from the baseline model for evaluation with all three datasets as described in the previous chapter. This chapter presents the results, followed by possible explanations for my findings.

5.1 Model performance comparison

Table 1 shows the error rates for both fine-tuned models and the baseline model for both noisy datasets and the unedited clean speech, as described in Chapter 4. From this table, it becomes clear immediately that fine-tuning of the models to improve noise robustness was successful, with both the WER and CER of the fine-tuned models decreasing strongly with a relative 48% and 60% respectively for the noisy test sets. For the clean test set, the performance difference between models is notably smaller with the noise-robust models still performing well.

Models	**Gym Noise**		**Public Noise**		**Clean**	
GymLibri	**9.9**	**4.5**	10.5	**4.7**	4.0	1.1
PublicLibri	10.9	5.1	**9.9**	**4.7**	4.0	1.1
Baseline Model	20.7	12.3	18.9	11.4	**3.4**	**1.0**

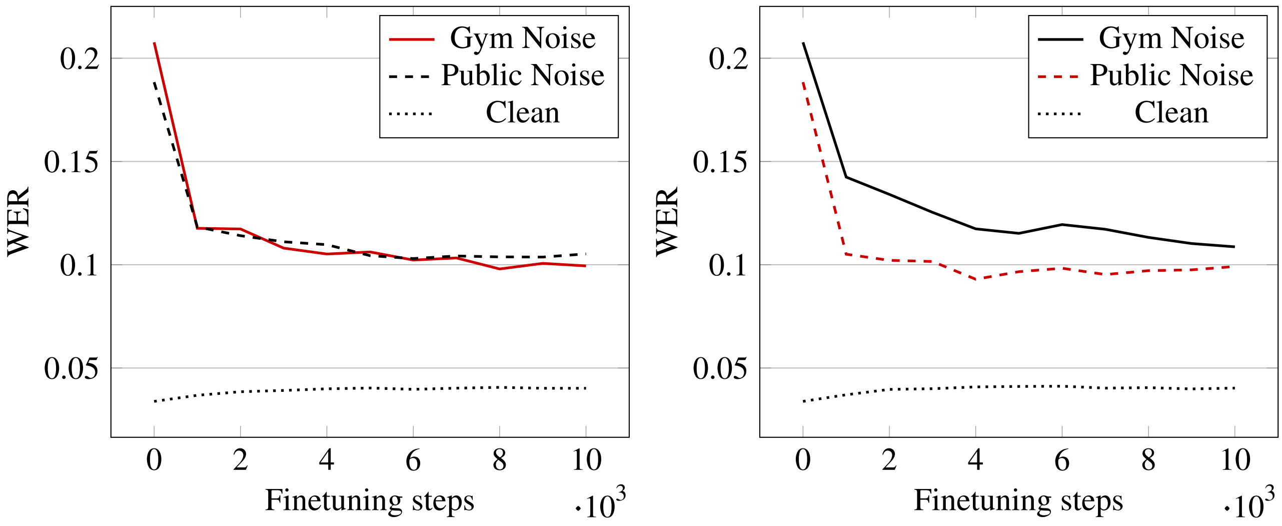
Table 1: Values of WER (left) and CER (right) for each model across the three datasets

Both fine-tuned models seem to have an advantage over the other for the test set corresponding to its training (the GymLibri model performs best for the Gym Noise test set, while the PublicLibri model performs better for the Public Noise test set). However, these differences are not large, and the models seem to transfer well to the other type of noise, indicating the presence of a general noise robustness. On average, the GymLibri model seems to outperform the PublicLibri model, with better transfer to the other dataset. Implications from these results are discussed in the next chapter.

5.2 Checking for overfitting

As is mentioned in section 2.3, overfitting is not expected to occur quickly for the GymLibri model, because of the great variety in the sound composition of the Gym Noise and the multi-condition training for the PublicLibri model. However, it remains important to make sure that the model is not overtrained. To monitor this, it is necessary to inspect the model performance throughout the fine-tuning process. The WERs and CERs in Table 1 are the result of fine-tuning both models for 10,000 steps, corresponding to 59 epochs. For every 1000 steps of training, a checkpoint was saved. In Figure 6, the performance for each checkpoint is depicted, showing the learning curve.

Figure 6a shows that already after 1000 steps of training, the WER drops quickly with a relative 37-42% for both noisy test sets. Throughout further training, the loss decreases quickly while the WER keeps going down at a slower pace. It stands out that performance of the two test sets stays very similar, even though only one type of noise is used for training. Performance for the clean test set rises slightly in the first 2000 steps and stays roughly constant during further fine-tuning. Figure 6b shows a similar scenario to Figure 6a, with the eye-catching difference being that the performance lines of the two noisy test sets are further apart. Whereas the performance data from the GymLibri model might suggest that the datasets are very similar, Figure 6b shows that there are in fact noteworthy differences between the two. Even though model performance stabilizes quickly for the Public Noise



(a) GymLibri model

(b) PublicLibri Model

Figure 6: Performance throughout training for all three datasets.

test set with the number of steps progressing, the WER for the Gym Noise test set steadily continues to decrease. For both models the WER is on average still going down slowly at 10000 steps for both models, there is no significant overfitting.

6 Discussion

6.1 Answering the research question

The research question that I aimed to answer with this thesis is stated in Chapter 3 as follows:

> "Does an end-to-end ASR model fine-tuned on noise recordings from a specific environment outperform a general noise-robust model for that specific environment?
> In particular, will this be the case for the specific noise environment of a climbing gym?"

Whereas Schlotterbeck et al. (2022) and Zhu et al. (2022) already reported that fine-tuning wav2vec2.0 on noisy data significantly improves noise robustness, it becomes clear that in this thesis the GymLibri model does outperform the PublicNoise model for the Gym Noise. This suggests that fine-tuning an end-to-end ASR model (wav2vec 2.0) on noise recordings from a specific environment does indeed improve performance more than general noise-robust models, when tested on audio from that particular noise environment. This is my primary hypothesis (H1). Nonetheless, one should bear in mind that such a broad claim cannot be made based on a comparison of only two models. To properly answer the general component of the research question, further comparison between models trained on different noisy datasets is required.

Furthermore, the earlier introduced secondary hypotheses are:

H2. The general noise-robust model will outperform the environment-specific model on other noises.

H3. Both noise-robust models will outperform the base (non-noise-robust) model for both noisy test sets.

H4. The base model will outperform both noise-robust models for clean test sets, but not with a large difference.

All the above can be accepted based on the results found. The PublicLibri model outperforms the GymLibri model for the Public Noise, which contains a larger variety of noises from different categories, suggesting that H2 can be accepted. Both fine-tuned models show a major improvement over the baseline model on the noisy test tests, confirming H3. Lastly, as stated in H4, the base model scores best for the clean test sets, but not by a large difference. As expected from the findings from Zhu et al. (2022), there is no large drop in performance of the fine-tuned models on clean speech.

It should be noted that the results suggest that the PublicLibri model is less noise-robust in general. Several possible explanations can be given for this, the first and most important being that noises were reused for different speech files and at different SNRs, instead of using only different noises. This choice was made to provide the best possible comparison with earlier research from Zhu et al. (2022). Even though data augmentation like this has been proven to be efficient for improving model performance (Sivasankaran et al., 2017), more original data might still be preferred for the model to learn a broader representation of noise. This puts the quality of comparison between the two training sets and models up for debate.

6.2 Limitations

This thesis knows multiple limitations which are discussed below. To support readers and raise awareness for these issues, the key words of each item are written in bold.

One limitation of this thesis is the **lack of significance testing** for difference in performance, which unfortunately could not be conducted because it did not fit in the time frame of this thesis. While reviewing literature in the field, I found that information on statistical tests is missing in many papers comparing performance of different models or for different datasets. However, in order to make well-supported solid claims about which model outperforms the other, a significance test is essential. I acknowledge this shortcoming in my research and would like to encourage others to pay attention to this topic and at least mention whether or not they conducted a statistical test. The word 'significant' often seems to be used in papers in a casual manner without mentioning any statistical test, preventing readers from knowing whether a difference is noticeable or actually statistically significant. I would like to mention a helpful tool that is used in several papers that did include statistical significance reports, called the NIST Speech Recognition Scoring Toolkit[8]. In this thesis, the difference of over 50% relative WER improvement is so apparent that it can be assumed safely that this is not by chance, but for smaller differences one must be cautious when drawing conclusions (e.g. for the differences in performance of the fine-tuned models on the noisy datasets, and for the differences in performance of all models on the clean set).

Another factor which makes the reported results less generalizable is that I chose to use splits from the **same speech corpus both for testing as well as for training** the model. This issue was addressed by Szymański et al. (2020), as explained in section 2.2. This choice was made to best compare with earlier research since it is common practice to split a corpus in three parts for training, validation and testing. Better would have been to use a test set from Mozilla's Common Voice[9], for example, or compare model performance on different corpora. Another large factor was the time available for this master's thesis, which did not allow to add too many layers of complexity like comparing performance on different speech datasets mixed with the different noise datasets that were used.

Another weakness of this thesis is that even though the final models are meant for real-world use with spontaneous speech, they are **trained on read speech**. This means that performance will be lower than the reported 9.9% WER for the GymLibri model when it would actually be employed in the climbing gym. As already mentioned in the introduction of this thesis, this remains a common problem in the speech recognition field (Szymański et al., 2020). Related to this, the setup of **mixing noise with clean speech as opposed to using direct recordings of speech in the noisy environment** is different from natural recordings, making it a suboptimal approach for making an ASR model as suited to the real world as possible. The reverberation that might occur when speaking in the climbing gym does not occur when mixing clean speech with the background noise. However, this would make the recording process very time-consuming and impractical. At busy moments, there is often little free space in the gym, and actively making recordings with speakers with a varied range of voices requires a lot of time and effort.

Even though the speech is not recorded in the most realistic way, the noise recorded inside the climbing gym has great variation. In essence, this is a good thing, since it means that all kinds of sounds and scenarios are well-represented and the dataset represents the complete sound setting in a realistic way. However, due to the difficult recording circumstances, unnecessary extra variation may have been introduced. Especially the **placement of the microphone** could have been more consistent, as this has great impact on the recordings. It is unknown how this has affected model performance, and if this variable had a positive or negative impact. It is probable that this extra variation has made the model more robust to general noises, but it might also have been too much variation for the model to learn the specific patterns of the gym noise environment.

[8] `https://github.com/usnistgov/SCTK`

[9] `https://commonvoice.mozilla.org/`

Amongst all noises, speech noise is included in the Gym Noise dataset. The main language spoken in the climbing gym from this study is **Dutch**, while the model is trained on English speech only. This choice was made, again, to better compare with other works. To my knowledge, the **influence of the language of background speech on ASR performance** has not been researched and is likely to be minimal at least for this thesis, because speech elements in the recorded noise were generally poorly intelligible. To create an ASR model that would actually be applicable for this gym however, it would thus have to be trained on Dutch instead of English.

6.3 Suggestions for future research

As becomes clear from this chapter, several suboptimal decisions were made with the reasoning of staying in line with the research field. Whereas it was not feasible for me to make changes in these topics while still reaching the goals of this thesis and providing good comparisons, the world of speech technology would greatly benefit from studies focusing on these topics and aiming to broaden and improve common research practices. I would like to encourage researchers working in the same field to critically think about the topics mentioned in the previous section, and not just follow earlier employed common methods.

Not for all issues raised above is it known how severe they influence the outcomes. For example, I am not aware of any research done on the difference between the methods of recording speech directly in a noisy environment as opposed to mixing clean speech with background noise. It would be a valuable contribution to fine-tune ASR models on these two types of data and compare their performance on both types of data. With the outcome of this, we could determine if these two methods are closely comparable and which is the best approach for ensuring ASR robustness. Another interesting topic of research is the influence of the language of babble noise or background speech on model performance. For example, when Mandarin babble noise is mixed with Dutch speech, does this influence model performance compared to when Dutch babble noise is mixed with Dutch speech? With the outcome of this, we could determine how relevant language elements are in background noise.

After reviewing the outcomes of this study and the topics above, there are still many questions that remain with many opportunities for research. Most important is the generalizability of my findings to other noise environments and models trained on other noisy datasets (both from specific noise environments and general noise collections). A relevant and interesting opportunity would be to evaluate our model on noisy data from another climbing gym and see how well the model generalizes to that similar but deviating noise environment.

Lastly, this study only presented and employed one relatively basic method of making an ASR model noise-robust (fine-tuning on noisy speech data). A reviewing study replicating and comparing different methods proposed in different papers would shed light on the question which is the most efficient. Since each study uses slightly different methods, it is difficult to compare models exactly from only reading the papers. An overarching study on the costs, complexity and effectiveness of each method would be of great value.

7 Conclusion

This study investigated the method of fine-tuning an end-to-end speech recognition model to achieve optimal noise robustness for one specific environment. While ASR models have achieved remarkable performance in the ideal conditions often presented in research, their performance is still often falling short in real-world settings. With indoor sport climbing quickly gaining popularity, but gyms often being situated in facilities with poor acoustics, the noisiness of the environments can be an obstacle for ASR users.

This thesis tested if specializing an ASR model to a climbing gym would be a solution to this problem for a fixed ASR application in the gym, or that any general noise-robust ASR model would give similar performance. Two wav2vec 2.0 models (pre-trained on an English LibriSpeech dataset of 960 hours) were fine-tuned on LibriSpeech datasets mixed with noise: one dataset was mixed with noisy recordings from the climbing gym, and the other with various types of noise from daily real-world situations. Both fine-tuned models outperformed the baseline model by an impressive 47-48% relative WER improvement for noisy speech. The model fine-tuned on Gym Noise outperformed the general noise-robust model for the gym noise with a relative 6% in terms of WER. Both fine-tuned models maintained high performance for clean speech. These results suggest that fine-tuning an end-to-end ASR model on noise recordings from a specific environment does indeed provide an advantage over general noise-robust models, when tested on audio from that particular noise environment.

With this finding, this thesis contributes not only to the field of noise-robust ASR, but also to the representation of sport climbing in literature and bridging the gap between these two areas. To my knowledge, this thesis is the first work bringing these two topics together by targeting speech technology, ASR specifically, in the context of a climbing gym environment. This contributes to creating an open and inclusive space where people are more aware of invisible disabilities and the obstacles that people with these disability experience. There are endless opportunities for speech technology implementations and even though there is still a long way to go to reach perfection, it is important to raise awareness of possible applications to people outside of the academic speech technology scene, to connect these worlds and pave the way for future applications. I encourage researchers to continue improving for all types of real-world scenarios and broaden the inclusivity and robustness of ASR systems.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer [computer program] (version 6.1.53).* Retrieved from `http://www.praat.org/`

Du, J., Wang, Q., Gao, T., Xu, Y., Dai, L.-R., & Lee, C.-H. (2014). Robust speech recognition with speech enhanced deep neural networks. In *Fifteenth annual conference of the international speech communication association.*

Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st acm international conference on multimedia* (pp. 411–412).

Higuchi, Y., Tawara, N., Ogawa, A., Iwata, T., Kobayashi, T., & Ogawa, T. (2021). Noise-robust attention learning for end-to-end speech recognition. In *2020 28th european signal processing conference (eusipco)* (pp. 311–315).

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., ... others (2021). Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846.*

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., ... Synnaeve, G. (2020). Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint arXiv:2010.11745.*

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech communication*, *22*(1), 1–15.

Maas, A., Le, Q. V., O'neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust asr.

NKBV (Royal Dutch Climbing and Mountaineering Federation). (n.d.). *Klimmen in nederland.* Retrieved 2023-07-29, from `https://nkbv.nl/kenniscentrum/klimmen-innederland.html`

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Prasad, A., Jyothi, P., & Velmurugan, R. (2021). An investigation of end-to-end models for robust speech recognition. In *Icassp 2021-2021 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6893–6897).

Schlotterbeck, D., Jiménez, A., Araya, R., Caballero, D., Uribe, P., & Van der Molen Moris, J. (2022). "teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data. In *International conference on artificial intelligence in education* (pp. 269–280).

Shrawankar, U., & Thakare, V. M. (2013). Adverse conditions and asr techniques for robust speech user interface. *arXiv preprint arXiv:1303.5515.*

Sivasankaran, S., Vincent, E., & Illina, I. (2017). Discriminative importance weighting of augmented training data for acoustic model training. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4885–4889).

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., ... Carmiel, Y. (2020). Wer we are and wer we think we are. *arXiv preprint arXiv:2010.03432.*

Taub, M. (2022). *Why we need to do more to support the deaf climbing community.* `https://www.climbing.com/people/supporting-deaf-climbers/.`

Van Segbroeck, M., & Narayanan, S. S. (2013). A robust frontend for asr: combining denoising, noise masking and feature normalization. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 7097–7101).

Wang, D., Shangguan, Y., Yang, H., Chuang, P., Zhou, J., Li, M., … Chandra, V. (2021). Noisy training improves e2e asr for the edge. *arXiv preprint arXiv:2107.04677*.

Wang, Y., & Gales, M. J. (2012). Speaker and noise factorization for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(7), 2149–2158.

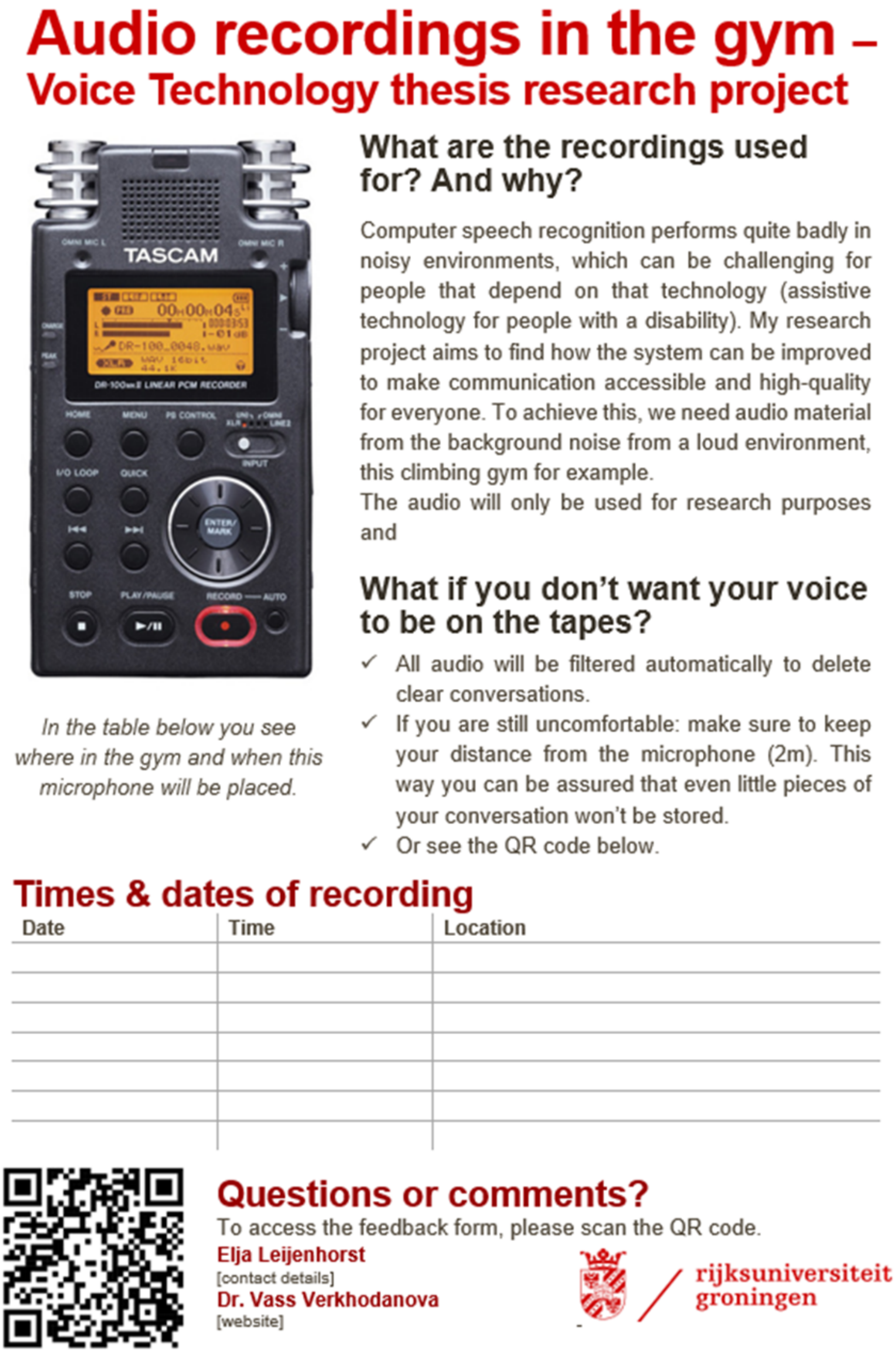
Wang, Y. Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 ieee workshop on automatic speech recognition and understanding (ieee cat. no. 03ex721)* (pp. 577–582).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wróbel, J., & Pietrusiak, D. (2021). Noise source identification in training facilities and gyms. *Applied Sciences*, *12*(1), 54.

Zhu, Q.-S., Zhang, J., Zhang, Z.-Q., Wu, M.-H., Fang, X., & Dai, L.-R. (2022). A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 3174–3178).

Appendix: Recording Information Flyer



Audio recordings in the gym – Voice Technology thesis research project

What are the recordings used for? And why?

Computer speech recognition performs quite badly in noisy environments, which can be challenging for people that depend on that technology (assistive technology for people with a disability). My research project aims to find how the system can be improved to make communication accessible and high-quality for everyone. To achieve this, we need audio material from the background noise from a loud environment, this climbing gym for example.

The audio will only be used for research purposes and

What if you don't want your voice to be on the tapes?

- ✓ All audio will be filtered automatically to delete clear conversations.
- ✓ If you are still uncomfortable: make sure to keep your distance from the microphone (2m). This way you can be assured that even little pieces of your conversation won't be stored.
- ✓ Or see the QR code below.

In the table below you see where in the gym and when this microphone will be placed.

Times & dates of recording

Date	Time	Location

Questions or comments?

To access the feedback form, please scan the QR code.

Elja Leijenhorst
[contact details]
Dr. Vass Verkhodanova
[website]

rijksuniversiteit
groningen

Figure 7: This flyer with information on the recording process was placed throughout the climbing gym to inform visitors. Table contents were filled in manually after printing and updated regularly.