University of Zurich[UZH]

Master's Thesis

for obtaining the academic degree

Master of Science

in Department of Informatics

Swiss German Speech-to-Text: Test and Improve the Performance of Models on Spontaneous Speech

Author: Karin Thommen

Student ID: 16-580-011

Advisor: Tanja Samardžić, PD Dr.

Supervisor: Martin Volk, Prof. Dr.

Department of Informatics

Faculty of Business, Economics and Informatics

Submission date: 01.08.2023

Abstract

Translators, voice recordings, and voice control are often pre-installed on mobile devices to make everyday life easier. However, Swiss German speakers must use Standard German or English when using speech recognition systems. The latest research shows that most of these systems are trained and evaluated on prepared speech. It remains an open question how these speech-to-text systems behave if they are applied to spontaneous speech, which consists of incomplete sentences, hesitations, and fillers. This can be summarised in the following research question: How does the performance of pre-trained speech models drop when fine-tuning on spontaneous speech compared to fine-tuning on prepared speech? Differences in speech styles lead to the assumption that performance drops when it comes to spontaneous speech. To assess the differences between prepared and spontaneous speech, two state-of-the-art pre-trained multilingual models were fine-tuned on the corresponding data. One is XLS-R developed by Facebook and proposed in 2022. Another model is Whisper by OpenAI, proposed in 2023. Thus, one main challenge is to make the models that are trained on two distinct speech styles comparable. Surprisingly, the results of both models disprove the hypothesis, as they perform better on spontaneous speech. Multiple improvement techniques were evaluated on their impact on the models. On the one hand, increasing the size of the data set significantly increases performance. However, one main issue in automatically transcribing Swiss German is finding the correct word boundaries. As many errors occur at the character level, it remains open which evaluation metric is the most appropriate for spontaneous speech and a low-resource language like Swiss German.

Zusammenfassung

Übersetzungsprogramme, Sprachsteuerung und Spracherkennung sind oft auf mobilen Geräten vorinstalliert, um den Alltag zu erleichtern. Diese Sprachsysteme werden laufend in verschiedenen Sprachen entwickelt, dennoch müssen Deutschschweizerinnen und Deutschschweizer in Hochdeutsch oder Englisch mit diesen Sprachsystemen kommunizieren. Die neueste Forschung zeigt, dass die meisten dieser Systeme auf sorgfältig vorbereitete Sätze trainiert wurden. Entsprechend bleibt offen, wie sich ein Modell verhält, wenn es auf spontane Diskussionen und Gespräche angewendet wird, welche sich durch Füllwörter und unvollständige Sätze auszeichnen. Dies kann in der folgenden Forschungsfrage zusammengefasst werden: Wie verändert sich die Leistung eines Sprachmodells, das auf spontane Sprache trainiert wurde im Vergleich zu einem Modell, das auf vorbereitete Sprache trainiert wurde? Der Unterschied in den beiden Sprachstilen lässt annehmen, dass die Qualität eines Modells sinkt, wenn es auf spontane Sprache angewendet wird. Um den Leistungsunterschied zwischen vorbereiteter und spontaner Sprache zu zeigen, wurden zwei Modelle aus der aktuellen Forschung als Basis verwendet und auf die beiden Sprachstile trainiert. Das Modell XLS-R wurde von Facebook entwickelt und im Jahr 2022 vorgestellt. Ein weiteres Modell ist Whisper von OpenAI aus dem Jahr 2023. Eine Herausforderung hierbei ist es, die Modelle für die beiden unterschiedlichen Sprachstile vergleichbar zu machen. Überraschenderweise widerlegen die Ergebnisse beider Modelle die Hypothese, da sie bei spontaner Sprache besser abschneiden. Mehrere Verbesserungsmethoden wurden hinsichtlich ihrer Auswirkungen auf die Modelle untersucht. Es hat sich ergeben, dass die Vergrößerung des Datensatzes die Leistung erheblich steigert. Ein Hauptproblem bei der automatischen Transkription von Schweizerdeutsch ist jedoch das Finden der richtigen Wortgrenzen. Da viele Fehler auf der Zeichenebene auftreten, bleibt offen, welche Bewertungsmetrik für spontane Sprache und eine ressourcenarme Sprache wie Schweizerdeutsch am besten geeignet ist.

Acknowledgement

First, a special and big thank you goes to my advisor Tanja Samardžić, who supported me greatly for six months while writing the thesis. I was able to contact her at any time with questions or points for discussion and received ongoing feedback. The close supervision of the thesis was very valuable, and I would like to thank her for that. Another thank you goes to Martin Volk, who, as a supervisor, made it possible to write this thesis at the Institute for Computational Linguistics. I also thank Iuliia Nigmatulina, who provided me with some of the data.

Second, I would like to thank the Swiss Association for Natural Language Processing (SwissNLP) for allowing me to use the SDS-200 data set for my work. I also thank Manuela Hürlimann, research associate at the ZHAW School of Engineering, who supported me in finding a topic and data. She informed me about the current research in the area of Swiss German speech-to-text that is being worked on in Switzerland and connected me with people who could further support me in my work. These include Jan Deriu, Claudio Paonessa, Yanick Schraner, and Manfred Vogel from ZHAW and FHNW.

I would also like to thank the institutions where I was able to complete my studies and this thesis: the University of Zurich (UZH), the Department of Informatics, and the Institute of Computational Linguistics.

Finally, I would like to thank the people who privately supported me in my studies and work. This includes not only my family, but also my partner, who helped me to correct this thesis and other papers during my studies.

Contents

List of Figures

List of Tables

List of Acronyms

ASR	Automatic Speech Recognition
CER	Character Error Rate
CSV	Comma-separated values
CTC	Connectionist Temporal Classification
MT	Machine Translation
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
STT	Speech-To-Text
TSV	Tab-separated values
UTF-8	Unicode Transformation Format (8-bit)
WER	Word Error Rate
XLS-R	Wav2Vec2-XLS-R-300M

1 Introduction

This master thesis deals with Swiss German speech-to-text models. The focus is on the automatic transcription of spontaneous speech and on comparing the performance of speech-to-text models for prepared and spontaneous speech. Many systems in recent research are trained with sentences that are carefully read by a native speaker. However, one main application of speech-to-text models is to transcribe natural conversations consisting of fillers and hesitations. These transcriptions could then be used to transcribe discussions in TV shows or videos on social media platforms to make them, for example, accessible to people with a hearing disorder. Another gap in speech-to-text research is the development of speech-to-text systems in low-resource languages, especially Swiss German.

In the scope of this thesis, spontaneous speech is defined as the speech style used in conversation (Tucker and Mukai, 2023). The opposite of spontaneous speech can be called prepared speech, which is carefully read speech, for example, reading sentences from a book or reading sentences from previously written notes. The terminology differs among different research papers and books. In this thesis, spontaneous speech will be used for a dynamic speech during an unprepared conversation where a person has no time to carefully plan the sentences he or she is about to say. Grammatical errors, uncompleted sentences, breaks, and filler words characterise this type of speech. On the contrary, the term prepared speech refers to speech in which the speaker reads out some neat sentences that are mostly grammatically correct, thought-through, and, therefore, less noisy.

1.1 Motivation

Language systems such as translators, voice recordings, voice control, and voice input are being developed in various languages. They are automatically installed on mobile devices and should make everyday life easier (Gabler et al., 2023). However, Swiss German speakers must still communicate with these systems in English or Standard German. Given that there are more than five million Swiss German speakers (Bundesamt für Statistik, 2022), however, there is a growing need for certain functions to also exist in Swiss German (Scherrer et al., 2019b). In addition, accessibility is important. Subtitles to television shows are essential for barrier-free daily life, especially for people with hearing disabilities. This requires transcription systems that automatically transcribe Swiss German discussions and insert the transcription as subtitles in Standard German.

The deficiency in developing language systems is not limited to Swiss German but is prevalent in many low-resource languages. As Magueresse et al. (2020) describe in their paper about low-resource languages, there are around 7000 languages worldwide, but today's research in natural language processing (NLP) mainly focuses on 20 languages. The problem is that, apart from these 20 languages, most are low-resource, meaning that they are less studied or taught (Magueresse et al., 2020). But already in Africa or India, about 2.5 billion people speak a low-resource language (Magueresse et al., 2020). Therefore, the relevance of language system development has economic and social aspects (Tsvetkov, 2017). Research in Swiss German speech-to-text has started but mostly focuses on prepared speech (Plüss et al., 2022).

1.2 Research Questions

Latest research shows that most automatic speech recognition models (ASR) are trained on carefully read speech data (Furui et al., 2005). As one main application area of ASR models is the recognition or transcription of speech during a natural conversation, the performance of this use case is of great interest. This leads to the following research question as the centre of this thesis:

> *How does the performance of pre-trained speech models drop when fine-tuning on spontaneous speech compared to fine-tuning on prepared speech?*

Spontaneous speech is very different from prepared speech. On the one hand, the way of collecting the data is different. On the other hand, these are different speech styles, as Chapter 2 shows. This leads to the assumption that performance drops when it comes to spontaneous speech compared to models trained on prepared speech.

1.3 Challenges

Research groups are currently developing language models that can handle Swiss German, as can be seen at Swiss conferences in the field of NLP[1]. Existing models are primarily trained on Standard German text sentences recorded in Swiss German Schraner et al. (2022). But this is not a real scenario of language usage. When speaking, people tend to restart sentences and correct themselves during a sentence or a word (Horii et al., 2022). Moreover, during discussions, people interrupt each other in the middle of a sentence. Traditional ASR models have problems handling this type of speech (Horii et al., 2022). However, spontaneous speech is important for speech technology (Tucker and Mukai, 2023). The goal of this thesis is to tackle the problem of Swiss German speech-to-text systems with a focus on spontaneous speech. To do this, the performance gap between prepared and spontaneous speech must first be estimated. There is a main challenge to making the models that are trained on two distinct data sets comparable. On the one hand, the challenge includes preprocessing the data in a similar way. On the other hand, the experimental settings of the two pre-trained models, XLS-R and Whisper, must be set as comparable as possible for fine-tuning.

[1] `https://www.swisstext.org/programme/`

1.4 Thesis Structure

The first chapter is the introduction and focuses on the motivation for the topic of the thesis and presents the research question. Chapter 2 introduces concepts on automatic speech recognition systems and speech styles. Moreover, the chapter summarises related work and describes the gap in which this thesis is anchored. In chapter 3, the data and the model architectures are described. In addition, the preprocessing and evaluation metrics are introduced. Chapter 4 presents the results of the research carried out in the scope of this thesis. This includes the performance of the models developed on the test set, as well as the training and validation loss and the word error rate during training. The fifth chapter deals with interpreting and discussing the results of Chapter 4. The last chapter concludes all the findings and insights on prepared and spontaneous speech and the two state-of-the-art model architectures used in the scope of the thesis.

2 Background

The chapter summarises the concepts and related work. The concepts are about language, dialects, and speech styles, focussing on Swiss German. The following is a description of how automatic speech recognition systems work. Furthermore, the section defines the terminology of spontaneous speech and prepared speech as used in this thesis and other related research. Chapter 2.2 outlines research in similar fields and illustrates the gap in which the thesis is anchored.

2.1 Concepts about Languages and Systems

The following paragraphs describe the concepts underlying this thesis. These concepts include languages and speech styles, fundamental concepts for explaining automatic speech recognition, and other related terms and theories. Moreover, terms are defined as they are used in other research and within the scope of this thesis. This includes an explanation of speech-to-text and automatic speech recognition and how ASR systems are trained. After explaining the technological background of speech-to-text, the terms prepared speech and spontaneous speech will be clarified. In research, various terms refer to different speech styles (Tucker and Mukai, 2023). For the sake of simplicity, spontaneous speech and prepared speech are used as opposing terms in this thesis.

Languages Switzerland has four official languages: German, French, Italian, and Romansh (Gsteiger and von Cranach, 2012). The language of the majority of speakers in Switzerland is German (Bundesamt für Statistik, 2022), especially Swiss German. More than five million people speak Swiss German, most of them living in Switzerland (Bundesamt für Statistik, 2022). Swiss German is a collection of German dialects that vary between regions and differ in phonetics, vocabulary, morphology, and syntax (Schraner et al., 2022). The main difference between Swiss dialects compared to dialects in other European countries is the width of the usage of the dialects (Scherrer et al., 2019b). Swiss dialects are used in different domains,

such as public speech or education (Scherrer et al., 2019b). Years ago, Swiss German dialects were mainly used for spoken communication (Scherrer et al., 2019b). With the advance of technologies and the development of digital communication tools, Swiss German is also used for written communication (Scherrer et al., 2019b). Swiss people also write text messages in their Swiss German dialect, but there is no standard orthography (Schraner et al., 2022). Differences in Swiss dialects and the lack of defined rules make it difficult to develop natural language processing systems in Swiss German (Schraner et al., 2022).

Speech Styles There exist not only different dialects but also different styles of speech. In addition to dialects, the different speech styles also complicate automatic speech recognition. The style of speech is defined through internal and external factors that depend on the situation, formality, and mood (Tucker and Mukai, 2023). For example, speech differs by gender, age, social class, and education (Scherrer et al., 2019a). This thesis focuses on two contrasting speech styles, namely, prepared and spontaneous speech, independent of other aspects. The terms will be defined later in this section.

Speech-to-Text / Automatic Speech Recognition Speech is used as communication between humans (Juang and Rabiner, 2005). With the advancement of technologies, the desire to automate communication processes with the help of modern technology has increased (Juang and Rabiner, 2005). Automatic Speech Recognition (ASR) is the process of transforming speech signals into a textual representation (Gabler et al., 2023). ASR systems must be strong, robust, and flexible to be used in daily life (Gabler et al., 2023). Traditionally, automatic speech recognition was based on statistical language modelling (Juang and Rabiner, 2005). Later, deep neural network-based hybrid modelling is becoming the standard in the development of automatic speech recognition systems (Li, 2021). The latest research is shifting to end-to-end models for automatic speech recognition (Li, 2021). There are different end-to-end techniques in the field of automatic speech recognition. One technique is connectionist temporal classification (CTC) to map speech input sequences into an output label sequence (Li, 2021). This technique is widely used for ASR. Another approach is the attention-based encoder-decoder model (Li, 2021). This model consists of an encoder network, an attention module, and a decoder network (Li, 2021). A third example of an end-to-end technique for ASR is the recurrent neural network (RNN) transducer that provides a natural way for speech recognition (Li, 2021). The RNN transducer can transform any input sequence into a finite discrete output sequence because RNNs can store and access information about sequences

over a longer period of time (Graves, 2012). For CTC, conditional independence is assumed, which is no longer used in the case of the RNN transducer (Li, 2021).

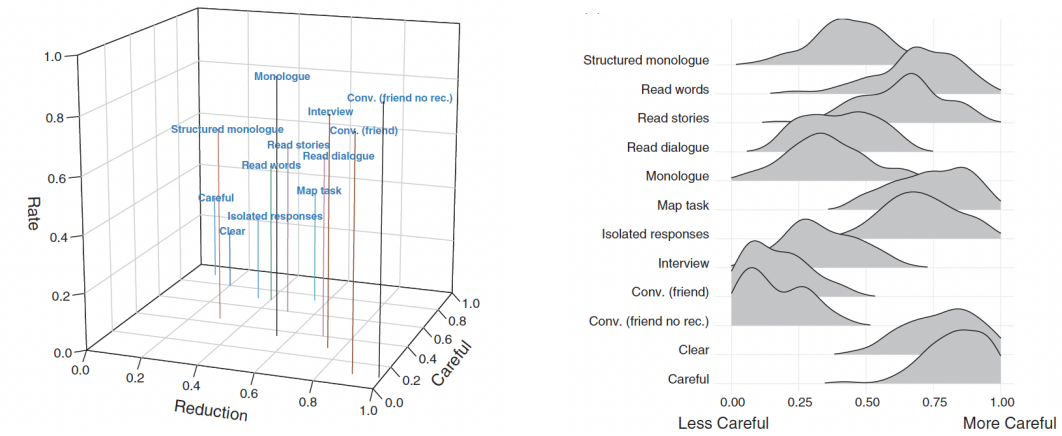
Training ASR systems Most ASR systems are trained with supervised machine learning, which requires training data consisting of labelled data (Gabler et al., 2023). In the case of speech recognition, there are recorded chunks of speech with the corresponding transcriptions (Gabler et al., 2023). In the case of prepared speech, this would be sentences and the complementary audio of a speaker who reads the sentences (Gabler et al., 2023). In the case of Swiss German, the audio can be recorded in different dialects and differ from the written sentence in word order and vocabulary. The training corpora for spontaneous and prepared speech differ in the way of collection. To collect the prepared speech data, various speakers record themselves while reading the given sentences (Gabler et al., 2023). As Gabler et al. (2023) describes, for spontaneous speech, annotators are needed that reconstruct the conversation and write a transcription manually. This leads to more noise in the training data, making the development of ASR systems more difficult (Gabler et al., 2023).

Swiss German and ASR Mobile devices are integrated into daily life, and computational power has increased in recent years (Gabler et al., 2023). Moreover, machine learning technology has also been advanced and further developed, and new possibilities for human and machine interactions have opened (Gabler et al., 2023). Language systems are developed for languages other than Swiss German (Scherrer et al., 2019b). Speech recognition works well primarily for languages such as German or English, but not for Swiss German (Plüss et al., 2021). The problem is that Swiss German is primarily a spoken language, and for informal texts, there is no standardised grammar (Plüss et al., 2021). This leads to the difficulty of developing automatic speech recognition systems (Schraner et al., 2022). Some main reasons are spelling ambiguities and the large size of the vocabulary (Schraner et al., 2022). An approach to these problems is to convert Swiss German speech to Standard German text (Schraner et al., 2022).

Spontaneous Speech As Tucker and Mukai (2023) shows, the terminology of the speech styles differs through research. In the scope of this thesis, the terms spontaneous speech and prepared speech will be used as opposites. First, the term spontaneous speech will be described in detail in similar research. Tucker and Mukai (2023) says that spontaneous speech is a speech style that refers to conversational, connected, casual, fast, natural, and vernacular speech. As Tucker and Mukai (2023)

summarises, the main characteristic of spontaneous speech is that it is not read. Spontaneous speech is harder for automatic speech recognition due to variation and noise (Gabler et al., 2023).

Prepared Speech The opposite of spontaneous speech is prepared speech. In this sub-chapter, the term will be described as used in the scope of this thesis and documented in other research. (Tucker and Mukai, 2023) does not use the term prepared speech explicitly, but describes the opposite of spontaneous speech as follows: careful, read, laboratory, scripted and formal speech. (Tucker and Mukai, 2023) emphasises that a political speech or an acted conversation can sound natural, but it does not mean that the speech is spontaneous. Prepared speech is essential in speech research (Tucker and Mukai, 2023).



(a) Multiple dimensions of speech (Tucker and Mukai, 2023)

(b) Range of carefulness of different speech styles (Tucker and Mukai, 2023)

Figure 2.1: Illustrations of speech variations and their variability and overlap within dimensions (Tucker and Mukai, 2023)

The variety of speech styles is important for research (Tucker and Mukai, 2023). It depends on the research goal on which speech style is the focus. But the different types of speech vary in various dimensions, as proposed by Tucker and Mukai (2023) in their definition of spontaneous speech. Tucker and Mukai (2023) show with the figure 2.1 some different speech styles that vary in carefulness with how a sentence is produced. Tucker and Mukai (2023) defines carefulness with the way speech is articulated. Careful speech is hyper-articulated, while conversational speech consists of more hesitations and shorter segments (Tucker and Mukai, 2023). Figure 2.1a shows a three-dimensional representation of speech styles. The rate describes the velocity, ranging from slow to fast speech (Tucker and Mukai, 2023). The reduction represents the degree of abbreviation of sentences and words. On the one hand, there are read words or read text where the speech style is careful. These are examples of

prepared speech. On the other hand, there are interviews or conversational speech, which are less careful speech styles and an example of spontaneous speech.

Performance Difference between Prepared and Spontaneous Speech One of the main differences between prepared and spontaneous speech is how to collect training data (Gabler et al., 2023). This leads to difficulty in applying models trained on prepared speech to spontaneous speech, as Gabler et al. (2023) shows in their analysis. There exists research on prepared and spontaneous speech in other languages that has shown that the performance of ASR on spontaneous speech is lower than on prepared speech Gabler et al. (2023). One reason is that speakers tend to pronounce more precisely if they read a sentence, which is the case for prepared speech Gabler et al. (2023). Spontaneous speech leads to higher degrees of segmental reduction and greater variation in speech rate, the usage of fillers, self-correction, or repetition (Gabler et al., 2023). These differences between speech styles are shown in Figure 2.1. Due to the difference in data collection, spontaneous speech transcriptions tend to be noisy (Gabler et al., 2023).

2.2 Related Work

The research and development of automatic speech recognition is advancing in many languages (Yadav and Sitaram, 2022), for example, English or German. Current speech-to-text or automatic speech recognition works on single clean sentences, also called prepared speech (Furui et al., 2005). But in reality, spoken communication is noisy (Gabler et al., 2023). People stop in the middle of a sentence and restart it or correct themselves during speaking (Horii et al., 2022). Moreover, this results in sentences that are not grammatically correct (Tucker and Mukai, 2023). In summary, research in spontaneous speech is not yet as advanced as in prepared speech. For this thesis, the SDS-200[1] corpus is used as a data set for the prepared speech. The data set collectors developed a baseline model when they introduced their data set and used the XLS-R model. For the XLS-R model with 1B parameters, they received a Word Error Rate (WER) of 21.7% on the validation set; in the test set, they reported a WER of 21.6% (Plüss et al., 2022). This data set and three others were used in another article to test the performance of different systems (Schraner et al., 2022). They report the performance of different ASR systems on four different data sets. All data sets consist of Swiss German audio and Standard German transcripts. Most of the data sets consist of prepared speech data. Although they tested different

[1] `https://swissnlp.org/datasets/`

systems and reported their performance on various dialects, they did not test the performance of the ASR systems on spontaneous speech. The gap here is that they do not show whether the systems work only on prepared speech data sets or whether they would also work if spontaneous speech is given as input. They mention free speech and dialogues in the conclusion of their paper as future work (Plüss et al., 2022).

2.3 Research Gap

In other languages, ASR is already better, even for spontaneous speech (Yadav and Sitaram, 2022). Although much work is currently being done on developing Swiss German Speech-to-Text models, research in this area is not as advanced as in other languages, as described in Chapter 2.2. In summary, there is a gap in Swiss German speech-to-text, especially in spontaneous speech. An example of the difference between prepared and spontaneous speech is shown in Furui et al. (2005). Horii et al. (2022) propose disfluency labelling as an approach to improve the performance of a spontaneous speech model. Horii et al. (2022) mention hesitations and fillers as one of the main problems of spontaneous speech and one of the reasons why the performance drops if spontaneous speech is used as data instead of prepared speech. They developed a baseline on prepared and spontaneous speech data to prove this. For prepared speech, they register a character error rate (CER) of 4.5-5.4% (Horii et al., 2022). For spontaneous speech, they report a CER of 16.0%. After applying their approach of disfluency labelling, Horii et al. (2022) achieved a CER of 10.3% for spontaneous speech and 3.8-4.5% for prepared speech.

On the one hand, they showed that the performance of a model is lower for spontaneous speech than for prepared speech (Horii et al., 2022). On the other hand, they proposed a way to improve performance with disfluency labelling (Horii et al., 2022). However, it is unclear whether these findings apply directly to Swiss German or other languages since the research was in Japanese.

Furui et al. (2005) showed in their research that the performance of automatic speech recognition systems drops from prepared to spontaneous speech. The goal is to prove this performance difference in these two speech styles in Swiss German and to investigate strategies to improve the quality of ASR systems in the case of spontaneous speech.

3 Methodology

The chapter 3 describes the approach of how the goal of the thesis was addressed. Moreover, the chapter presents the data used to train the model and the different pre-trained model architectures used in the scope of the thesis. The section 3.4 is about the evaluation metrics, the data preparation, and the experimental settings used to fine-tune speech-to-text models. This includes the further preprocessing of the data, the baseline models, and some improvement strategies.

3.1 Approach

The research question of the thesis is to find out how automatic speech recognition systems perform on prepared and spontaneous speech. In addition, some possible approaches should be tested to improve the models on spontaneous speech data. To compare the performance of models on the two speech styles and to prove promising ideas of improvement for Swiss German, an own baseline has to be developed. For the baseline, existing models from previous research will be used. To compare prepared and spontaneous speech, two separate models were developed. One model will be fine-tuned on prepared speech, while another model will serve as a baseline for spontaneous speech as it is fine-tuned on another data set. Subsequently, the goal is to improve the results of the baseline model by testing different strategies suggested by recent research in the field of automated speech recognition.

The baseline on spontaneous speech can then be used to evaluate how automatic speech recognition can be improved with a focus on Swiss German. As a reference and starting point, recent research by Plüss et al. (2022) and Schraner et al. (2022) was used. They used Wav2Vec2-XLS-R models as a basis for their research on Swiss German speech-to-text. Since the models and data sets of the named papers are resource consuming to train, the baseline was developed on the basis of a smaller model with fewer parameters and smaller data sets.

Huggingface provides different versions of the Wav2Vec2 XLS-R model[1]. As recent research by Schraner et al. (2022) uses the large model with 1B parameters, a smaller version was used for the thesis. As a consequence, the smaller model will produce worse results than the current state-of-the-art. However, it is easier to apply some improvement strategies in the last step since it requires fewer resources during fine-tuning. Nevertheless, it can be used to compare the performance of models trained on prepared or spontaneous speech data and to determine what is needed to improve the performance of spontaneous speech. Future work could apply the results and findings to more extensive state-of-the-art models.

The last step is to test different approaches to improve the spontaneous speech model. These are approaches like the adaption of the training data or the preprocessing. Moreover, the impact of different hyperparameters can be shown in this part.

3.2 Data

For this thesis, two types of data sets are needed – one for each speech style. The first data set consists of annotated prepared speech audio to establish a baseline and for later comparisons. The second data set consists of audio transcriptions containing spontaneous speech. This data set will be used to establish a second baseline to compare spontaneous and prepared speech. In addition, the data set will be used to test improvement strategies for ASR models. Both data sets consist of an audio file and a corresponding textual transcription of the audio for each sample in the set. The main difference between the data sets is how the data is collected and how the corresponding text is produced. For the prepared speech data set, the text was given first, and the audio was produced for the given text. This means that people read the given text in their own dialect and record an audio. The data set containing spontaneous speech samples was collected the other way around. The audio was first given as the collectors of the data set retrieved the audio from the 'Schawinski' TV show. Then, a transcription of the audio was produced in retrospect. Regarding a single sample from the data set, the textual representation is called 'transcription' for both types of data set, in the scope of this thesis.

[1] `https://huggingface.co/docs/transformers/model_doc/xlsr_wav2vec2`

3.2.1 Prepared Speech: SDS-200

Schweizer Dialektsammlung (SDS-200) was used as the prepared speech data set. The data set was previously collected and is available through SwissNLP[2] and was proposed in Plüss et al. (2022). The reason for choosing this data set was a recommendation of the researchers who worked on Plüss et al. (2022) and Schraner et al. (2022). As Schraner et al. (2022) also used XLS-R, they recommended using SDS-200 in the scope of the thesis with respect to resources. Using more data would be time consuming, and insights on the performance difference between prepared and spontaneous speech can be shown using a smaller data set. The development of a distributable model achieved by using more data and larger models will be left for future work. The goal of the thesis is to prove the performance difference between prepared speech and spontaneous speech and not to achieve the best results by using more computational power.

The SDS-200 data set consists of 200 hours of recorded sentences. Each sample has information about the dialect, age, and gender of the 4000 speakers who contributed to the audio collection. The speakers recorded sentences that were prompted in their dialect on a publicly open web tool. Furthermore, speakers could validate other recordings to increase the quality of the audio collection. As described at the beginning of Chapter 3.2, the data set is an example of prepared speech, as the data are collected by letting people read some written sentences. The resulting audios are carefully read sentences, grammatically correct, read without hesitations or fillers, and recorded with little noise. The paper by Plüss et al. (2022) presents more information on the collection of the data set and the performance of some models trained on the data set.

The whole corpus consists of 152'251 sentences, but not all samples were used to train the baseline in the scope of this thesis. The speakers validated the audio of some recordings by other participants. Therefore, some clips were labelled as 'not valid'. A clip is considered valid if two or more participants identified the audio as a correct Swiss German representation of the original written Standard German sentence. A clip is not valid (value **clip_is_valid**: False) if the clip was rejected by more than two other participants. There are also clips that were not validated because there were not enough votes on the clip to decide whether it is a correct representation or not. The analysis of the data set shows that most of the clips do not have enough votes.

[2] `https://swissnlp.org/datasets/`

The following list shows the number of clips for the different types of validation statuses:

- Number of validated clips: 44'272
- Number of not validated clips: 6314
- Number of clips with not enough votes: 101'665

The SDS-200 data set consists of three splits: Training, validation, and test data. The table 3.1 shows the size of the splits of the data set.

Data Set	Size (number of items)
Training Set	135271
Validation Set	3638
Test Set	3636

Table 3.1: The distribution of items in the different data set splits of the SDS-200 data set presented in numbers.

Validity and Quality The analysis of the data set and the size of the splits show that non-valid audios were deleted from the data set splits. For the training set, audio files with not enough votes are still included. There are only validated clips for the validation and test data, as the table 3.2 below presents.

Validity	Training Set	Validation Set	Test Set
Number of validated clips	36973	3638	3636
Number of not validated clips	0	0	0
Number of clips with not enough votes	98298	0	0

Table 3.2: Distribution of validated and not validated clips in the different data set splits of the SDS-200 data set presented in numbers.

The original data set also reports a value called **user_mean_clip_quality**. This is a measurement of the mean quality ratio of the clip in intervals 0 to 1. An analysis of these values shows that for the training set the mean over all samples for the training set is 0.91, with a minimum value of 0.049 and a maximum value of 1.0. For the validation set, the mean is 0.93, the minimum value 0.76, and the maximum 1.0. For the test set, the mean for all samples is 0.91, the minimum value is 0.7, and the maximum is 1.0.

Dialects Table 3.3 shows the distribution of the different dialects for the three splits of the data set. The dialect of the canton Zurich is the one with the most audio files for the training set. For the validation and test set, there are more samples of the Bernese dialect than the Zurich dialect. This presents another difference compared to the data set consisting of spontaneous speech, which only consists of samples recorded in the dialect spoken in Zurich.

Canton	Training Set	Validation Set	Test Set
ZH	45068	208	277
AG	15624	151	55
BE	14381	441	425
VS	11486	126	127
SG	7546	76	41
GR	3228	0	9
TG	2438	42	192
FR	2372	10	55
LU	2232	196	118
SO	2161	10	24
BS	1948	0	23
BL	1503	10	56
SZ	1112	58	10
SH	904	61	0
AI	842	9	0
ZG	691	56	18
AR	549	90	39
OW	486	10	19
NW	184	13	0
GL	122	19	9
UR	110	0	51
JU	19	0	0

Table 3.3: Distribution of dialects in the data set splits of the SDS-200 data set presented in numbers.

Gender Table 3.4 shows the gender distribution in the data sets. Especially for the validation and test set, for most of the data, the gender is unknown. In all data sets, there are more known male audio clips than female speakers.

Gender	Training Set	Validation Set	Test Set
Male	47.7% (64570)	19.5% (710)	9.1% (334)
Female	19.5% (26371)	7.2% (265)	6.9% (253)
NaN	32.7% (44280)	73.2% (2663)	83.8% (3049)

Table 3.4: Distribution of gender in the different data set splits of the SDS-200 data set presented in numbers and percentage.

Age Table 3.5 is about the age distribution in the three splits of the data. For the training set, the age categories from the twenties to the fifties are well-distributed. For the validation set, we have similar numbers of samples for the 20s, 30s, and 40s. Also, for the test set, there are more samples for the 30s and 50s and only a few or no samples for the other categories.

Age	Training Set	Validation Set	Test Set
Teens	3564	75	0
Twenties	17773	236	111
Thirties	23051	282	202
Forties	18998	232	47
Fifties	21847	51	203
Sixties	5435	99	24
Seventies	287	0	0
Eighties	35	0	0

Table 3.5: Distribution of age in the different data set splits of the SDS-200 data set presented in numbers.

Transcriptions In the process of the data set collection, Standard German sentences were prompted to native speakers. The analysis of these sentences shows that the Standard German sentences have a mean length of 50 characters in all data set splits. The minimum length lies between 15-20 characters, and the maximum length is around 100-120 characters.

Example The example 3.1 shows how an entry in the SDS-200 data set is composed. Each value containing the word 'id' refers to an identification number of the clip, the sentence, or the participant who recorded the audio. The source of the sentences could be from the Swiss newspaper 'Tamedia' (value: **tamedia_sentences**) or is retrieved from the German Common Voice[3] (value: **cv_sentences**).

(3.1) **Example-Annotation**

clip_id	127972
clip_path	8587421a-f201-4e8a-95a9-e3d3f2473f5e/ 0f97ad53f8c704aa2d9b0200bcf5718bacae8b 58128106d0072432fde948a747.mp3
sentence	'Dies wird inzwischen durch Forschungsergebnisse widerlegt.'
clip_created_at	2021-08-25 08:20:01
clip_is_valid	NaN
sentence_id	0f97ad53f8c704aa2d9b0200bcf5718bacae8b 58128106d0072432fde948a747
sentence_source	cv_sentences
client_id	8587421a-f201-4e8a-95a9-e3d3f2473f5e
zipcode	5000
canton	AG
user_mean_clip_quality	0.9011
clip_n_votes_correct	0
clip_n_votes_false	0
clip_n_times_reported	0
sentence_n_times_reported	0
age	fourties
gender	male
duration	4.176
user_sentences	1038
continuous_client_index	NaN

More examples, including some audios, are available on the SwissNLP website[4]. A small sample corpus can be downloaded there as well.

[3] Ardila et al. (2020) and `https://commonvoice.mozilla.org/de`

[4] `https://swissnlp.org/datasets/`

3.2.2 Data on Spontaneous Speech: Schawinski

The data containing spontaneous speech were already collected in the scope of other research, but the data set is unpublished. As described in Chapter 2, spontaneous speech data is collected differently from the prepared speech data. In this case, a television talk show was transcribed into text and divided into chunks. The television talk show was the 'Schawisnki' talk show, where the host, Roger Schawinski, interviews well-known Swiss people.

Speakers and Setting For the described 'Schawinski' data set, 30 minutes of six talk shows were transcribed. The six people and Roger Schawinski all speak the dialect of the canton of Zurich. Transcripts from discussions of the following people are part of the data set: Fredi Hafner, Allan Guggenbühl, Christoph Mörgeli, Doris Fiala, Jacqueline Badran, Dieter Meier, and Ursula Schäppi.

The interviews were divided into small chunks and each chunk is annotated with its transcription and some information about the audio. Information about the chunks is saved in a CSV file. There are in total 4836 chunks with a mean length of the transcriptions of 45 characters. Table 3.6 presents some numbers about the data set.

Information	Numbers
Size	4836
Speech-in-speech	2268
no-relevant-speech	133
Mean transcription length	45 characters
Minimum transcription length	2 characters
Maximum transcription length	224 characters
Length data set	4836

Table 3.6: Information about the Schawinski data set.

Transcriptions The CSV file contains an identifier, an annotation, a speaker identification number, the duration of the segment, and information if the audio contains **speech-in-speech** or **no-relevant-speech**. **Speech-in-speech** means that the interviewee and Roger Schawinski are talking simultaneously. **No-relevant-speech** means that there is, for example, only music or some other fillers. Both values contain a 0 or 1 to indicate whether the tag applies to the audio chunk.

Example Example 3.2 provides a sample line from the CSV file to show how the data set is composed.

(3.2) **Example-Annotation**

utt_id	Badran_Schawinski_13-05-2013_SPK0-Badran_Schawinski_13-05-2013-0002
transcription	*'[music] da isch di eerschti taakschou vo de wuche froit m/ mi dass si au hüt aabig debii sind bi miir isch d schagglin badraan ässphee nazionaalräätin [breath_mouth_noise]'*
speaker_id	Badran_Schawinski_13-05-2013_SPK0
duration	6.39
speech_in_speech	0
no_relevant_speech	0

Special Annotations Compared to the SDS-200 data set, the Schawinski data set is not only collected differently. The data set consists of text annotations of audio and some labels or remarks on hesitations, fillers, and other non-speech content. These special annotations can be part of a sentence and appear in the middle, at the end, or at the beginning of a sentence. But some audios exist where the transcription consists only of a special annotation. These annotations are marked as **no_relevant_speech** in the data set.

A list of all special annotations that mark no-relevant-speech is given by table 3.7.

Special Annotation	Description
[breath_mouth_noise]	noises like breathing / smacking
[speech-in-speech]	more than one person speaking
[music]	background music
[laughter]	laughing vocals
[noise]	isolated noise
[speech-in-noise]	noise during speech
$	hesitations
/	pauses and interruptions

Table 3.7: Special annotations which are part of the transcription of the Schawinski talk shows.

As the examples of both data sets show, the transcription language differs for the two data sets. The prepared speech data set consists of Standard German annotations. The spontaneous speech audio files were transcribed in Swiss German according to an adaptation of the Dieth rules in written Swiss German (Dieth and Schriftkommission, 1938).

3.3 Model Architecture

The research question deals with the comparison of a model's performance, whether it is trained on spontaneous or prepared speech. Although the data set is smaller to save computational resources during training, some current state-of-the-art models are used to answer the research question to ensure comparability with the latest research.

A state-of-the-art model is Wav2Vec2-XLS-R developed by Facebook and proposed in 2022. Another model is Whisper by OpenAI proposed in 2023. As both models are used in recent research in the field of automatic speech recognition, their performance was measured on spontaneous and prepared speech data to compare their behaviour.

3.3.1 Architecture and Background of Wav2Vec2-XLS-R

In the case of Wav2Vec2-XLS-R, the version with 300M parameters was used in the scope of this thesis[5]. XLS-R is a large-scale multilingual pre-trained model for learning cross-lingual speech representation (Babu et al., 2022). The basis of the model is wav2vec 2.0 which is a framework that masks speech input in the latent space (Baevski et al., 2020). Babu et al. (2022) propose XLS-R as a model trained on 436k hours of unlabelled speech in 128 languages. The model can be fine-tuned on downstream tasks for automatic speech recognition and was trained in multiple versions up to 2B parameters (Babu et al., 2022). The model is trained self-supervised, which means that most of the training data is unlabelled (Babu et al., 2022). The data are called unlabelled because the data sets consist of hours of audio from different sources with no textual transcription of the audio. One source is 50K hours of books read in 8 languages, 7K hours of single sentences read in 60 languages, 6.6K hours of YouTube speech in 107 languages, 1K hours of phone conversations in 17 languages, and 372K hours of Parliament speech in 23 languages (Babu et al., 2022). XLS-R is called a cross-lingual model because it is trained in multiple languages, and there does not exist a separate model for each language (Babu et al., 2022). As Babu et al. (2022) summarise in their paper, cross-lingual speech representation models provide similar performance compared to monolingual models in the field of speech recognition.

As mentioned, the basis of XLS-R is wav2vec 2.0 which is a current state-of-the-art model for automatic speech recognition that addresses the problem of the availability of labelled data (Baevski et al., 2020). In many fields, producing a large amount

[5] `https://huggingface.co/facebook/wav2vec2-xls-r-300m`

of labelled data that can be used for supervised training is resource consuming and time consuming, but speech recognition systems require a large amount of labelled data, which is not available for all languages, especially not for low-resource languages (Baevski et al., 2020). Baevski et al. (2020) argue that human language acquisition works differently and that this process can be adapted to ASR systems. Children learn languages by listening to adults and learning a representation of speech (Baevski et al., 2020). Therefore, they suggest using self-supervised learning in which a model is trained on unlabelled data, and labelled data is only used to fine-tune (Baevski et al., 2020).

XLS-R is built on the wav2vec 2.0 model architecture. In wav2vec 2.0 a convolutional feature encoder maps the audio to a speech representation (Babu et al., 2022). The speech representations are input to a Transformer, which outputs context representations (Baevski et al., 2020). XLS-R is optimised with Adam, the learning rate has a polynomial decay to zero after 32K warm-up steps and the training audios were cropped to a length of 20 seconds (Babu et al., 2022). After training, the model is fine-tuned by adding a linear layer to predict the output vocabulary for speech recognition tasks (Babu et al., 2022).

There exist multiple versions of the model with different numbers of parameters. Training with the larger models would require more computational resources. Therefore, the small version with 300M parameters was chosen to compare spontaneous and prepared speech.

3.3.2 Architecture and Background of Whisper

Whisper[6] is a recently proposed state-of-the-art model that can be used for speech recognition tasks. Whisper is a pre-trained model for ASR and exists in multiple versions, while the small multilingual version was used in the scope of this thesis. This version consists of 244M parameters. Thus, comparability with the XLS-R model with 300M parameters is ensured. Whisper was trained with AdamW optimiser and the learning rate decays to zero after a warm-up for 2048 steps, and the audio was cropped to a length of 30 seconds (Radford et al., 2023).

The model was trained on 680K hours of labelled speech data and was proposed by Radford et al. (2023). They refer to the insights of Baevski et al. (2020) and criticise that unsupervised training has some limitations in its usefulness since it needs fine-tuning to perform some specific tasks such as speech recognition. Their

[6] `https://huggingface.co/openai/whisper-small`

suggestion is scaling supervised pre-training. For data processing, they train Whisper without significant standardisation and rely on the ability of seq-2-seq models to map between samples and annotation (Radford et al., 2023). They used a data set consisting of pairs of audio and transcriptions extracted from different domains. The audio and transcription pairs were retrieved from the World Wide Web (Radford et al., 2023). The goal was to collect audio from different settings: different recording setups, different speakers, and different languages. Moreover, compared to XLS-R, Radford et al. (2023) did not only look for clean audio, but aimed for different audio qualities to increase the robustness of the model. Since they retrieved audio and transcriptions from the Internet, one problem during data collection was machine-generated transcription (Radford et al., 2023). Therefore, they had to filter the audio where the transcription tends to be machine-generated by another ASR system. Radford et al. (2023) mention some cues that help to detect machine-generated texts. One cue is punctuation, which is hard to detect from audio, so in machine-generated texts, there are no or only a few commas and missing complex punctuation such as exclamation points or question marks (Radford et al., 2023). Another indication is if the transcription is all uppercase or all lowercase (Radford et al., 2023). Together with the matching transcription, the audio segments were fed to the model for training after resampling the audio to 16kHz (Radford et al., 2023). They used an encoder-decoder transformer as proposed by Vaswani et al. (2017).

3.4 Methods

The next section describes the methods used to develop models to compare their performance if they are applied to prepared and spontaneous speech data. First, the evaluation metrics are presented. Then, all methods related to data preprocessing are described. The last sections are about the experimental settings of the developed models and the improvement strategies.

3.4.1 Evaluation Metrics

There exist multiple evaluation metrics to measure the performance of models in the field of automatic speech recognition or speech-to-text tasks, as seen in the research proposed in Chapter 2. These metrics include the word error rate and character error rate, as well as the BLEU score.

Word Error Rate The word error rate was used as a metric to evaluate the performance of the models. There exists a Huggingface library for different evaluation metrics. This library was loaded to measure the performance of the models. The word error rate (WER)[7] is a metric to evaluate the performance of automatic speech recognition systems and calculates the difference between the reference transcription and the prediction (Morris et al., 2004). Compared to the character error rate, it measures performance at the word level and not at the character level. The words in the reference transcription are aligned with the predicted word sequence to calculate what types of mutation are needed to restore the reference sentences (Morris, 2002). Before calculating the number of correct words (hits, H), the number of substitutions (S), deletions (D) and insertions (I) must be counted (Morris, 2002). All these numbers are added together in a formula:

$$WER = 100 \cdot \frac{S+D+I}{S+D+H}$$

In this case, the word error rate is displayed in percentage as it is multiplied by 100. The lower the word error rate, the better the performance of the speech system. The word error rate has no upper bound, as Morris (2002) describes. If the model starts hallucinating and inserting random words into a sentence, the predicted sentence becomes longer than the reference sentence, and the number of words inserted increases. This results in a word error rate greater than 100% (Morris, 2002).

Character Error Rate The calculation of the character error rate was also performed using the Huggingface library[8]. Character error rate (CER) is another metric to measure the performance of automatic speech recognition systems Wang et al. (2013). Compared to the WER, it measures the difference between the reference transcription and the prediction at the character level instead of the word level. Therefore, deletions, insertions, hits, and substitutions are counted on the character level and again inserted into the following formula:

$$CER = 100 \cdot \frac{S+D+I}{S+D+H}$$

The character error rate could be interpreted similarly to the word error rate. The lower the number, the better the performance. The CER could exceed 100% if there are many insertions. Although many languages, for example German and Swiss German, are word-based, the character error rate is a common metric to measure errors in speech recognition for final evaluations of systems (Wang et al., 2013).

[7] `https://huggingface.co/spaces/evaluate-metric/wer`

[8] `https://huggingface.co/spaces/evaluate-metric/cer`

BLEU Score The BLEU score is another metric to measure the quality of a predicted text compared to the gold standard, but was only marginally used to evaluate the models for this thesis. The BLEU score can also be loaded using the Huggingface library[9].

3.4.2 Data Preparation

The following chapter describes the preprocessing of the data sets. The basis for preprocessing is audio files and CSV or TSV files containing some information about the data, as the original data sets were described in Section 3.2. The corresponding files can be found on GitHub in the 'Preprocessing' folder[10].

Data can be loaded efficiently if the data set is stored as a Huggingface data set and could be loaded using a Hugginface Hub repository[11]. **Datasets**[12] is another library available on Huggingface. The library can be used to preprocess audio for other NLP tasks. If the data are saved as a Huggingface data set, the data can be easily accessed and filtered (Huggingface, 2023b). Data and information are stored in Apache arrow format[13]. Therefore, large data sets can also be processed efficiently, without loading all data into memory (Huggingface, 2023c). Moreover, data sets are stored in a Huggingface Hub repository to save and load them during data preprocessing (Huggingface, 2023c). The arrow format stores data in a columnar memory layout (Huggingface, 2023c). Data sets can be stored in an on-disk cache that allows for a fast lookup of the data and is not memory consuming (Huggingface, 2023c). There exist multiple built-in functions to convert other data types into a data set (Huggingface, 2023c).

The Huggingface data sets were a solution to the slow processing of the audio files while using a pandas data frame to store the CSV data files and the information about the data. Each audio file had to be searched and accessed individually in the directories on the disk, which led to a memory overflow and increased processing time. Working with a Huggingface data set required a separate preprocessing step to store the audio files and the corresponding information in Apache arrow format and to upload them to the Huggingface Hub repository.

There exists an option to build a data set from scratch using the original data. It

[9] https://huggingface.co/spaces/evaluate-metric/bleu

[10] https://github.com/KarinTho/master-thesis-STT/tree/main/Preprocessing

[11] https://huggingface.co/docs/hub/repositories

[12] https://huggingface.co/docs/datasets/index

[13] https://huggingface.co/docs/datasets/about_arrow

is possible to create an audio data set with the local files with the AudioFolder builder, which is recommended if there are several thousand audio files, or to write a loading script. Since the SDS-200 data set consists of around 150'000 audio files, the AudioFolder was an appropriate solution. To load audio files into a data set with AudioFolder, the directory containing the audio has to be structured correctly. There must be a CSV metadata file named **metadata.csv**. This file has a row per audio file, with the first column containing the path to the file with the filename and the transcription of the audio in the second column. All other columns can be used to store other important information. In the case of the SDS-200 data set, information about the canton of the audio file and the duration of the audio is kept. For the Schawinski data set, only the duration is kept as additional information. At the same directory level as the metadata file, there must be a **data** folder where the audio files are stored. Huggingface data sets are optimised to store all splits (train, test, validation) in one data set. The structure is similar to a dictionary with the keys 'train', 'validation', and 'test'. Each key has another dictionary as a value that contains the 'audio'-data with the path and an array that describes the audio features, the sampling rate, and the transcription. In the case of the SDS-200, there are additional keys for the canton and the duration of the audio. SDS-200 collectors already split audio files into three parts: train, validation, and test. The information about which audio file belongs to which split was originally stored in a separate TSV file, and all audio files were stored in 4000 different folders, independent of the split to which they belong. Therefore, after creating **metadata.csv**, all files must be sorted so that all audio from the same split is stored in the corresponding folder.

- data/train/ - all training audios
- data/test/ - all test audios
- data/valid/ - all validation audios

After sorting all files according to the above structure and producing the metadata file, the AudioFolder can load all audio files. The data are converted to a parquet file to store the information in arrow format. As a last step, the parquet files are pushed to the Huggingface Hub repository, from where they can be loaded and used for training and testing.

An example shows how the first entry of the train split of the final preprocessed prepared speech data set looks like:

(3.3) **dataset['train'][0]**

{'audio': {'path': '09966c7743291ccf1129c8136143bf5a6132947fe352795bc6d5456a3afeb4de.mp3',
'array': array([0.00000000e+00, 0.00000000e+00, 0.00000000e+00, ..., 1.58690691e-05, -6.36559753e-06, -1.80013558e-05]),
'sampling_rate': 32000 },
'transcription': 'Dadurch wird auch der Lebensraum von vielen Tier- und Pflanzenarten zerstört.',
'canton': None,
'duration': 6.732}

The entries for spontaneous speech have the same structure.

Preparation of the Prepared Speech Data Set (SDS-200) The prepared speech data set is converted to a Huggingface data set as described above, with the splits kept as defined in the original data set.

Preparation of the Spontaneous Speech Data Set (Schawinski) The spontaneous speech data set originally was not divided into training, validation and test sets. Therefore, before storing all the data in a Huggingface data set, the splits have to be defined. There exist two preprocessed versions of the spontaneous speech data set. The first version is split as it was, with 20% testing, 80% training, with 20% for validation, with shuffled chunks among all speakers. Moreover, the first version does not contain any transcripts that contain a **speech-in-speech** or **no-relevant-speech** tag. As this split is not a natural setting for spontaneous speech, the second version is more appropriate for training a model. To keep the option of testing a model on longer and more coherent audio, chunks that appear after another should be kept in the same data set split, especially for the test split. Therefore, half of a discussion between a male and a female speaker and half of a discussion between two male speakers are extracted for validation and the test set. All other speakers are part of the training data. Contrary to Version 1, the two tags **no-relevant-speech** and **speech-in-speech** tags are not removed from the transcripts in this version of the data set. Thus, the special annotation can be used later.

3.4.3 Experimental Settings

The section on the setup of the training process describes all the default settings that are used to train the baseline models. It is important to mention that in some cases, the optimal settings that the developers suggested to use to fine-tune the pre-trained models are overwritten due to better comparability. The goal is to establish comparable baseline models by fine-tuning XLS-R and Whisper with prepared and spontaneous speech data.

Regardless of the type of model or the type of training data, the same parameters are set. Table 3.8 shows the hyperparameters used for XLS-R and Whisper for both types of speech. Especially during the development of the improvement strategies, some specific parameters are changed. If any parameter differs from the default values presented in Table 3.8, it is explicitly stated when a model and its results are introduced. The explanations of the hyperparameters can be found in the Glossary.

Training parameters	
per_device_train_batch_size	16
per_device_eval_batch_size	8
gradient_accumulation_steps	2
total_train_batch_size	32
learning_rate	0.0003
dropout	0
optimizer	adamw_hf
adam_beta	0.9 / 0.999
adam_epsilon	1.00E-08
lr_scheduler_warmup_steps	100
lr_scheduler_type	linear
save_steps	400
max_steps	4000

Table 3.8: Default training parameters for all models.

Baseline Model Wav2Vec2 XLS-R The first two baseline models are fine-tuned on the XLS-R model. There exists an example notebook that includes a tutorial on how to fine-tune an STT model in XLS-R[14]. This tutorial served as a starting point for the development of the baseline.

As a first step, the audio data set is loaded and, in the case of the prepared speech, the information about the canton is removed as it is no longer needed. This includes, for example, punctuation, square and round brackets, or mathematical operators.

[14] `https://huggingface.co/blog/fine-tune-xlsr-wav2vec2`

The deletion of special characters differs between the data sets due to a preceding analysis of the characters that appear. In the case of the spontaneous speech data, all special annotations are also deleted to make it comparable to the prepared speech data, as there exist no special annotations. Depending on the size of the data set, the data set has to be restricted to samples with a duration shorter than or equal to six seconds to avoid memory problems. Therefore, for the prepared speech data, there is always a filter to delete all files longer than six seconds from the data set. The spontaneous speech data set is much smaller and, therefore, memory overflow is not a problem. In all cases, the data gets shuffled at one point during preprocessing.

XLS-R takes a vocabulary file for its tokeniser. Therefore, all characters have to be extracted from the data set, and an enumerated vocabulary dictionary will be made from the three data set splits. In addition, unknown [UNK] tokens as well as padding [PAD] tokens are added. The last step is to save the vocabulary to a JSON file for later usage. The tokeniser, feature extractor, and processor can be loaded as the vocabulary file is prepared. The processor is a wrapper for the tokeniser and feature extractor. Similarly to the pre-trained model, the mentioned parts are based on wav2vec2. Since XLS-R is pre-trained on audios of 16kHz, the audios used for fine-tuning should be resampled as well. As soon as the audio is at the correct sampling rate, the input values should be loaded from the audio files. This happens with the wav2vec2 processor, which normalises the data.

The training process is based on the parameters presented in table 3.8. The option of early stopping is not implemented, as the goal is to train all models for the same number of steps to compare them.

Baseline Model Whisper Whisper differs in the way it prepares the data for training. In contrast to XLS-R, Whisper does not need a vocabulary file for its tokeniser. For multilingual fine-tuning, a language token can be added. Section 3.4.4 will discuss more about the language token. As for XLS-R, there also exists a notebook on Huggingface providing an example of how Whisper can be fine-tuned for ASR[15]. Similarly to preprocessing for the XLS-R fine-tuning, the data are also re-sampled to 16kHz in the case of the Whisper model. The data set is also shuffled, and the processor computed the input features from the input audio array, similar to the XLS-R version. The processor is again a wrapper for the tokeniser and the feature extractor.

Important to mention is that Whisper does not take audio longer than 30 seconds.

[15] `https://huggingface.co/blog/fine-tune-whisper`

This issue is not a problem, since for the prepared speech data, all audio longer than six seconds are deleted anyways to avoid a memory overflow. For spontaneous speech, there exists no audio longer than 30 seconds after preprocessing.

Data Set Sizes after Preprocessing After removing special characters and special annotations from the data sets, empty transcripts are also deleted. The two data sets are hugely different in size. Therefore, to obtain better comparability, the size of the prepared speech data set has to be reduced to a number similar to that of the spontaneous speech data. To investigate the impact of the size of the data set, some models should also be trained with a larger subset of the data set. For XLS-R all samples with a duration below six seconds are part of the data set. Due to a memory issue, for Whisper, only samples below five seconds are part of the subset. As a result, different types of preprocessed data sets are used to train the different models. For a better overview, the subsets are named and listed in this chapter. The results section will refer to these specific data sets.

	Small Prepared Speech	**Small Spontaneous Speech**	**Large Prepared Speech XLS-R**	**Large Prepared Speech Whisper**
train	1900	1857	135271	84771
validation	300	294	3638	1824
test	3636	393	3636	1603

Table 3.9: Data set sizes after preprocessing.

3.4.4 Improvement Strategies

This chapter describes different improvement strategies that were applied to the baseline models to find out how the performance of the spontaneous speech models could be improved.

Specify the Language The Whisper tokeniser does not take a customised vocabulary file as XLS-R does. However, a language token can be added to the processor and the tokeniser. Since the transcription of the prepared speech data set is in Standard German, the German language token is added to the processor and tokeniser. The transcription in the spontaneous speech data set is in Swiss German and not Standard German. Therefore, the language token can only be helpful in the case of the prepared speech data set. But for comparability, the German language token is also added to a version of the Whisper model that is trained on spontaneous speech.

Consider Punctuation for Whisper Whisper can predict punctuation. To make the baseline models comparable, punctuation and other special characters are deleted from the data sets during preprocessing. At a later point, a Whisper model is trained without removing punctuation from the data set to assess whether punctuation has an impact on the quality of the transcription. There is no punctuation by default in the spontaneous speech data set as there are no complete sentences. Therefore, a method is added to the preprocessing to add some punctuation at the end of a transcription to treat them as full sentences. It is important to mention that punctuation is not visible in any output of the model, as a normalisation step is applied with the tokeniser before calculating the word and character error rate. The normalisation deletes all special characters, converts the text into lowercase letters and applies other simple preprocessing steps to the ground-truth transcription and the prediction of the model. Radford et al. (2023) is reasoning for normalisation as the standardisation of different text forms to avoid a higher word error rate due to formatting or punctuation differences.

Disfluency Labelling In section 2 a strategy was proposed by Horii et al. (2022) that should increase the performance of models that are trained on spontaneous speech. Horii et al. (2022) describes fillers and hesitation as one of the main differences between prepared and spontaneous speech and one of the main problems while developing an ASR system for spontaneous speech. They suggested disfluency labelling as a possible solution to improve an ASR system and proved it in Japanese. The spontaneous speech data set has similar tags in the transcription that mark hesitations or fillers. There are '/' symbols to mark interruptions or pauses or '$' for hesitations. Moreover, 'laughter' and 'music' or 'breath mouth noise' are tagged in the transcriptions. Similarly to punctuation, special characters are not visible in the output of the model due to the normalisation step described by Radford et al. (2023). Therefore, only results on disfluency labels containing words can be reported in Chapter 4.

4 Results

This section shows the results of all the models developed based on the methods described in Chapter 3. The section is divided into six parts. The first part presents a comparison of prepared and spontaneous speech in the case of the XLS-R model and the Whisper model in a comparable version. These models are all marked with the term 'comparable' since the data was processed, and the hyperparameters of the models were set as similar as possible to draw conclusions on similar settings. The models in their comparable versions provide a baseline for further enhancements and experiments to assess the performance in various scenarios. The second section shows the impact of the size of the data set. To make the models in the first section comparable, the size of the data set was reduced. Therefore, the impact of the reduction is presented by comparing XLS-R and Whisper on a larger subset of the prepared speech data set. Section 4.3 provides the results of fine-tuned Whisper models with some default parameters, as suggested by the Whisper developers. The fourth section examines whether a model trained on prepared speech can be applied to spontaneous speech and vice versa. The following chapter then applies the results of various improvement strategies for spontaneous speech to the Whisper model. The last chapter presents the results of zero-shot learning of XLS-R and Whisper for comparison. For each model, the results are first presented and then interpreted. A concluding interpretation of all models, including a comparison of the different versions, is given in Chapter 5.

During the training process, training and evaluation losses are reported, as well as the word error rate in the validation set. To evaluate the performance of the models on an unseen data set, the models are saved after the training process and loaded again to be evaluated on the test split of the data set. When optimising the final versions of the models, different fine-tuned versions of the XLS-R and Whisper models are created. In the following chapters, the training parameters are presented only for the corresponding final model version.

For each model presented in the results section, some example predictions are provided to analyse the output of the model in detail and to see where the model produces some errors. For each data set, a reference sentence is chosen where the

prediction is presented for every model listed in this chapter. Hence, the output of the models can be compared based on this reference sentence. The examples always consist of the ground-truth transcription given by the data set and the prediction, which is the output of the model when the corresponding audio is given as an input. Moreover, the word and character error rate and the BLEU score are reported for the specific example.

The examples are always structured similarly to ensure better comparability and an overview of the different models. Each set of examples presents four different predictions. For the first prediction, a sentence was chosen from each data set. These sentences are shown as a first example for all models. The ground-truth for the first example of the prepared speech data set is the following sentence:

'die spanische verfassung erlaubt die abspaltung einer region nicht'

The ground-truth for the first example of spontaneous speech was another sentence since the test data of the data sets is different, and no sentences exist in both data sets.

'hät er min reschpäkt wil daas isch gaar käs chliises ämtli das isch es'

These two examples are illustrative of the speech style they represent. The example from the prepared speech data set is a complete sentence, while the spontaneous speech example is not. Additionally, each set of examples has three other sample predictions. The first sentence is always the same reference sentence, while the second one shows the sentence from the test set with the lowest word and character error rate. The third example is a prediction with a higher word error rate but a lower character error rate. The last prediction in the set of examples is always the prediction in the test set with the highest word error rate.

The predictions of the models on the test set are saved in CSV files. These files can be accessed on GitHub[1]. The corresponding testing script is also available on GitHub[2]. An overview of the models used in the thesis and their corresponding coding files can be found in the appendix A. The overview of the evaluation files is also part of the appendix.

[1] `https://github.com/KarinTho/master-thesis-STT/tree/main/Evaluation`

[2] `https://github.com/KarinTho/master-thesis-STT/blob/main/testing.ipynb`

4.1 Results of the Comparable Models

This section presents the results of the comparable version of the XLS-R and the Whisper model. For each data set, a separate model is fine-tuned. The parameters for the training process are set equally for all models to make them comparable. As the data set size has an impact on a model's performance, the sizes of the training sets are kept similar. The spontaneous speech data set consists of around 1900 samples after preprocessing. Therefore, for comparability, the training set of the prepared speech data is restricted to 1900 as well, containing only audio samples below a duration of six seconds to avoid a memory problem and to keep it comparable to any other versions that will be introduced later in Chapter 4. These models use the hyperparameters as described in Section 3.4.3 and serve as a baseline for later modifications.

4.1.1 Wav2Vec2-XLS-R

The first section presents the results of the XLS-R model developed by Facebook. As described in chapter 3, the pre-trained model 'facebook/wav2vec2-XLSR-300m' with 300M parameters is used to fine-tune. Two models are fine-tuned, one with prepared speech as training data and the other with spontaneous speech data. The models were then evaluated on the test set. As Table 4.1 shows, the model trained on spontaneous speech performs better in the test set than the model trained on prepared speech.

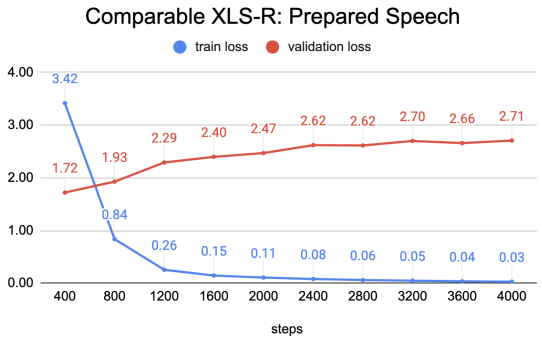
Comparable XLS-R			
Prepared Speech		**Spontaneous Speech**	
WER	79.95%	**WER**	60.89%
CER	34.48%	**CER**	16.74%
BLEU	0.054	**BLEU**	0.178

Table 4.1: The results of the prepared speech and spontaneous speech XLS-R models on the corresponding test set.

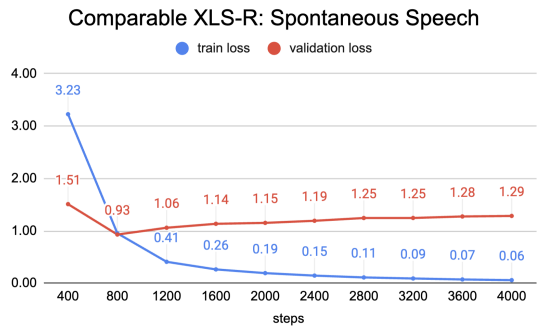
Each model is trained for 4000 steps, with a checkpoint every 400 steps. At each checkpoint, the training loss, the validation loss, and the word error rate gets calculated and reported.

Figure 4.1 shows the loss reported during the training process. In both cases, the training loss drops and converges to zero during the 4000 training steps. For the prepared speech version of the comparable XLS-R, the validation loss increases from the beginning, while the validation loss drops for the spontaneous speech model

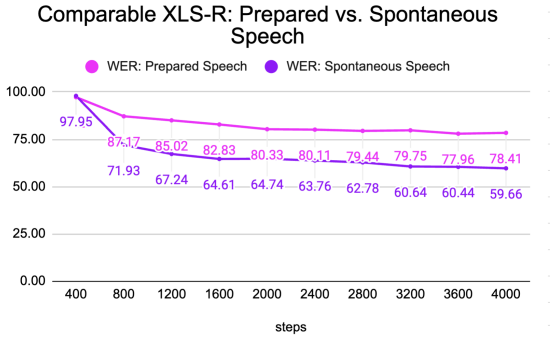
between the first two checkpoints. The third plot shows the development of the word error rates of both XLS-R models over all 4000 training steps. Both start with a word error rate of around 97-98% and decrease further with increasing number of steps. For both models, the curve flattens, but the spontaneous speech model reaches a significantly lower word error rate (59.66%). The table version of the plots can be found in the appendix.



(a) Train and validation loss of the comparable XLS-R model, fine-tuned on prepared speech data.



(b) Train and validation loss of the comparable XLS-R model, fine-tuned on spontaneous speech data.



(c) Word error rate of the prepared and spontaneous speech version of the comparable XLS-R models. The unit of the word error rate is displayed in percentage.

Figure 4.1: Training report of the comparable XLS-R model.

Examples 4.1 to 4.4 show some predictions of XLS-R if fine-tuned and tested on prepared speech. As described at the beginning of the chapter 4, the first example shows the prediction on the chosen reference sentence. The examples show that the model seems to have difficulties in detecting word boundaries, which leads to a strange composition of words such as, for example, 'einn den ereregion' in Example 4.1. Moreover, numbers are a problem, as seen in example 4.4. In this case, the model mixes written numbers with digits.

Comparable Model XLS-R – Trained on Prepared Speech

(4.1) **WER: 55.5% | CER: 15.15% | BLEU: 0.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'die spanischeven fassung erlaubt die abspaltung einn den ereregion nicht'*

(4.2) **WER: 0% | CER: 0% | BLEU: 1.0**
Transcription: *'doch in der zwischenzeit hat sich viel verändert'*
Prediction: *'doch in der zwischenzeit hat sich viel verändert'*

(4.3) **WER: 40% | CER: 3.33% | BLEU: 0.508**
Transcription: *'insofern ist das nicht störend'*
Prediction: *'in sofern ist das nicht störend'*

(4.4) **WER: 180.0% | CER: 103.57% | BLEU: 0.0**
Transcription: *'dies ist der artikel 546text'*
Prediction: *'das dart artikel fün 4r seix frogen zd det'*

The examples 4.5 to 4.8 show some sample predictions of the XLS-R model trained on spontaneous speech. Example 4.5 and 4.7 show that one problem of the model is to predict a blank space between words. In Example 4.8, a similar issue is recognisable. Moreover, it struggled to understand and transcribe the end of the sentence.

Comparable Model XLS-R – Trained on Spontaneous Speech

(4.5) **WER: 57.14% | CER: 10.0% | BLEU: 0.0**
Transcription: *'hät er min reschpäkt wil daas isch gaar käs chliises ämtli das isch es'*
Prediction: *'hät er minreschpäkt will daas isch gaarkäs s chliis es aämtli das isch es'*

(4.6) **WER: 0% | CER: 0% | BLEU: 1.0**
Transcription: *'mir chönd ois äfacht nüme erlaube'*
Prediction: *'mir chönd ois äfacht nüme erlaube'*

(4.7) **WER: 28.57% | CER: 4.05% | BLEU: 0.57**
Transcription: *'und dänn het mer seer vill chöne guetmache und es isch äfach wüürklich soo'*
Prediction: *'und dänn het mer seer vill chöne guet mache und es isch äfach würklich so'*

(4.8) **WER: 200.0% | CER: 24.0% | BLEU: 0.0**
Transcription: *'zwäihundertsächzgmilioone schwiizerfranke choschte'*
Prediction: *'zswäihunderzsächtz gmilioone schwiizer frald cha ch'*

All results of the XLS-R model show that it can predict umlauts, which is not the case for Whisper, as seen in the next chapter.

4.1.2 Whisper

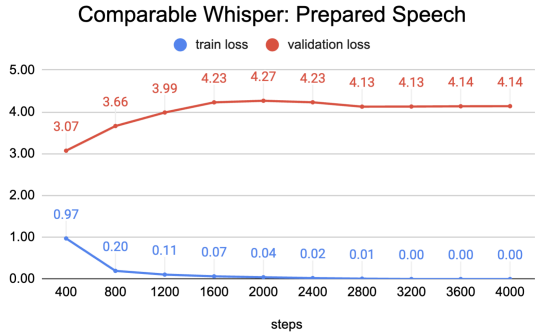
This section presents the results of the pre-trained model 'whisper-small' by OpenAI if it was fine-tuned on the prepared speech or spontaneous speech data. Similarly to the wav2vec2 XLS-R, Whisper was fine-tuned for 4000 steps with 400 steps between two checkpoints.

Comparable Whisper			
Prepared Speech		**Spontaneous Speech**	
WER	86.49%	**WER**	67.02%
CER	63.80%	**CER**	33%
BLEU	0.036	**BLEU**	0.156

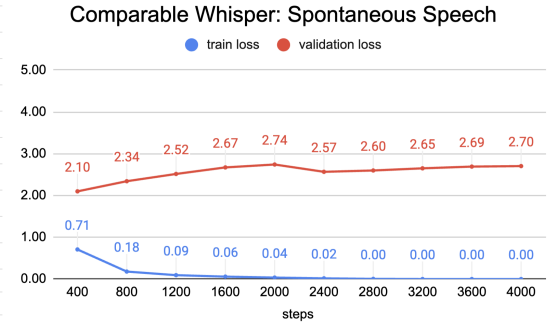
Table 4.2: The results of the prepared speech and spontaneous speech Whisper models on the corresponding test set.

Table 4.2 presents the results on the test set of the comparable Whisper model trained on prepared speech and the model trained on spontaneous speech. The spontaneous speech model performs significantly better than the prepared speech model. This can be seen in all evaluation metrics: the word and character error rates and the BLEU score.

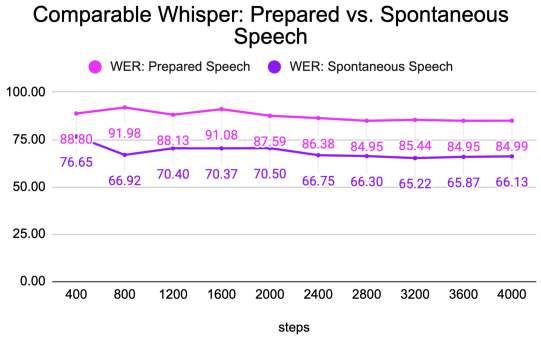
Figure 4.2 presents the loss of training and validation of the Whisper model trained on prepared speech and the one trained on spontaneous speech. The loss behaves similarly in both plots, but for the spontaneous speech model, the loss is overall lower. In both cases, the curve flattens after around 2000 steps. The word error rate in figure 4.2c behaves similarly for both speech styles, as it is correspondingly lower for spontaneous speech. After about 2000 steps, both models stop improving as the curve of the word error rate has also flattened. After 4000 steps, the prepared speech version reports a word error rate of 84.99%, and the spontaneous speech version has a word error rate of 66.13%. These results are significantly lower than those for the XLS-R model.



(a) Train and validation loss of the comparable Whisper, fine-tuned on prepared speech data.



(b) Train and validation loss of the comparable Whisper, fine-tuned on spontaneous speech data.



(c) Word Error Rate of the prepared and spontaneous speech version of the comparable Whisper. The unit of the word error rate is displayed in percentage.

Figure 4.2: Training report of the comparable Whisper model.

The sentences in the examples 4.9 to 4.12 show some predictions made by the model. The ground-truth and the prediction differ massively in the example with the worst results. With some extraordinary review of the audio and the predicted data, it was ensured that this was not an error in the data set or a mismatch between the ground-truth and the prediction. Sometimes whole words are missing, or two words are written as one. Noticeably, there is no word error rate below 11.1%. For the character error rate, the lowest value is 4.0%. The best BLEU score is 0.88.

It is important to mention that all Whisper models do not show umlauts in the output as a normalising step is applied. The reason is the normalisation step described in section 3.4.4, which is intended for English and removes all punctuation, makes everything lowercase, and replaces umlauts with the letters ‘aou’. As the normalisation step is added to the first versions of the Whisper models, it is kept for the other versions to make everything comparable.

Comparable Model Whisper – Trained on Prepared Speech

(4.9) **WER: 88.89% | CER: 56.06% | BLEU: 0.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'ist spanische verf in der lebt poststen viertig'*

(4.10) **WER: 11.1% | CER: 7.02% | BLEU: 0.88**
Transcription: *'uri war gestern fur eine stellungnahme nicht zu erreichen'*
Prediction: *'war gestern fur eine stellungnahme nicht zu erreichen'*

(4.11) **WER: 28.57% | CER: 8.89% | BLEU: 0.643**
Transcription: *'solche firmen mussen so mehr steuern bezahlen'*
Prediction: *'solche firmen mussen so mehr steuerzahlen'*

(4.12) **WER: 180.0% | CER: 141.38% | BLEU: 0.0**
Transcription: *'auch ich wurde etwas kriegen'*
Prediction: *'ist er vom verband fur mehrere spiele gesperrt worden'*

The comparable Whisper model tends to transcribe only a few words as in example 4.13. Often, words seem random, and the dependence on the ground-truth is not recognisable. The predictions in these cases consist of around one to three words.

(4.13) **WER: 100.0% | CER: 87.5% | BLEU: 0.0**
Transcription: *'rund ein dutzend falle betrafen die stadt zurich'*
Prediction: *'betroffen'*

Example 4.14 shows that for the model, it is difficult to reconstruct the sentence from Swiss German to Standard German if the sentence structure is different. The grammar and word order in Swiss German can differ from Standard German. If the prediction is aligned with a Swiss German transcription, it can be seen that the model does not rearrange the words as they should be in Standard German but tries to transcribe them as they are in spoken Swiss German.

(4.14) **WER: 100.0% | CER: 86.96% | BLEU: 0.0**
Transcription: *'ich brach in tranen aus'*
Prediction: *'aus bingen die ausbrachen'*
Swiss German: *'ich bin in Trääne uusbroche'*

Whisper achieves better results if fine-tuned in spontaneous speech, as seen in Table 4.2 and Figure 4.2. Some sample predictions are presented in the examples 4.15 to 4.18. Again, finding the word boundaries seems to be a problem for the model. Moreover, the difference between the word and the character error rate is significant, as there are more small errors on character level than a whole misunderstood word. An example is the prediction in 4.17, which looks primarily correct but has an 'e'

that occurs in the wrong position. The result is a high word error rate but a low character error rate.

Comparable Model Whisper – Trained on Spontaneous Speech

(4.15) **WER: 50.0% | CER: 17.14% | BLEU: 0.0**
Transcription: *'hat er min reschpakt wil daas isch gaar kas chliises amtli das isch es'*
Prediction: *'hat er miere schpakt i daas isch parargkas chliese s andli das isch es'*

(4.16) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'das mer uber daas debattiert und das hat'*
Prediction: *'das mer uber daas debattiert und das hat'*

(4.17) **WER: 33.3% | CER: 4.65% | BLEU: 0.508**
Transcription: *'di eerschte hundertfufzgtuusig franke vo de'*
Prediction: *'die eerscht hundertfufzgtuusig franke vo de'*

(4.18) **WER: 200.0% | CER: 50.0% | BLEU: 0.0**
Transcription: *'labe'*
Prediction: *'lab a'*

4.2 Impact of the Data Set Size

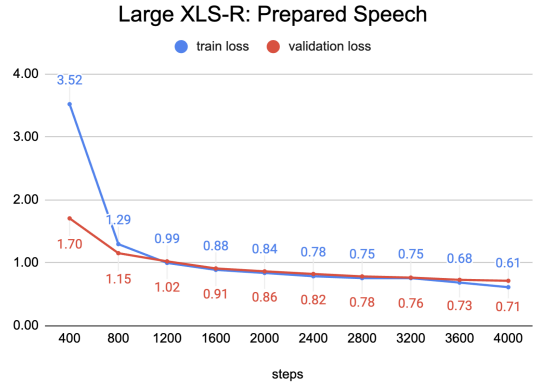
In order to ensure a comparison, the size of the prepared speech data set is reduced. The size of the data set can influence the performance of a model and its likelihood of overfitting. Therefore, it is important to report the results of fine-tuning XLS-R and Whisper in larger data sets as well. Therefore, XLS-R and Whisper are trained on a larger subset of the prepared speech data set. The other settings of the parameters are kept similar to compare the results. For XLS-R, all samples with a duration below six seconds are part of the data set, which means that XLS-R uses a larger subset than Whisper. Due to a memory issue, for Whisper, only samples below five seconds are part of the subset, namely the large prepared speech Whisper data set. The detailed numbers of the data sets can be found in Chapter 3.4.3.

It is important to mention that in the case of the large Whisper model, the German language token is added to the Whisper tokeniser. The comparable Whisper version is once trained without any language token and once with the German token added to the tokeniser. This leads to slightly better results in the case of prepared speech data. Therefore, the German token is kept for the large Whisper version.

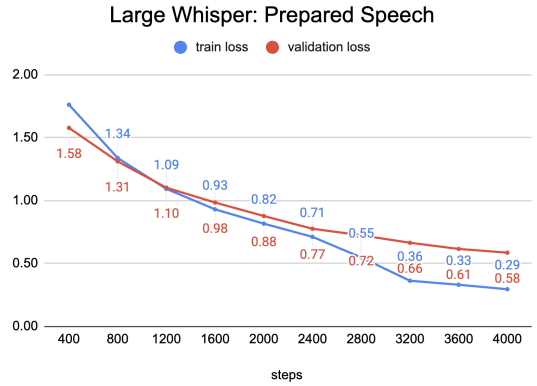
Large XLS-R		**Large Whisper**	
Prepared Speech		**Prepared Speech**	
WER	36.03%	**WER**	64.45%
CER	16.46%	**CER**	45.19%
BLEU	0.422	**BLEU**	0.233

Table 4.3: The impact of the data set size on XLS-R and Whisper if they are trained on a larger subset of the prepared speech data than the corresponding comparable version.

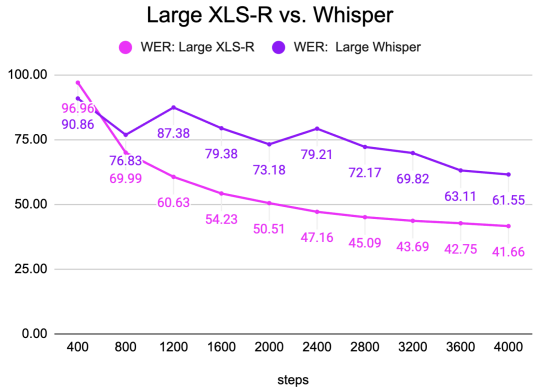
The results of the two larger versions are presented in Table 4.3. The comparable XLS-R model, which was trained on a smaller data set, achieves a word error rate of 79.95%, a character error rate of 34.48% and a BLEU score of 0.054 in the prepared speech test set. If XLS-R is fine-tuned on more data, the word error rate drops to 36.03% with a character error rate of 16.46% and a BLEU score of 0.422. This is significantly lower compared to the version with the smaller data set.



(a) Train and validation loss of the large XLS-R, fine-tuned on prepared speech data.



(b) Train and validation loss of the large Whisper, fine-tuned on prepared speech data.



(c) Word error rates of XLS-R and Whisper if fine-tuned on a larger subset of the prepared speech data. The unit of the word error rate is displayed in percentage.

Figure 4.3: Training report of the large XLS-R and Whisper models.

Some sample predictions of the large XLS-R model are presented in examples 4.19 to 4.22. The reference sentence in 4.19 shows that the number of errors decreased, as the only problem in the predicted sentence is a repetition of the word 'einer'. The example 4.22 is the sentence in the test set with the worst prediction. As the prediction is significantly different from the transcription, the audio is manually checked. It turns out that the transcription is correct regarding the content, but originally the speaker was articulating the words predicted by the model. Therefore, it is open to discussion how it could be handled that there often exist multiple correct transcriptions of a spoken sentence.

Large XLS-R Model – Trained on Prepared Speech

(4.19) **WER: 11.11% | CER: 7.576% | BLEU: 0.658**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht '*
Prediction: *'die spanische verfassung erlaubt die abspaltung eine einer region nicht'*

(4.20) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'der sieg war keine grosse sache '*
Prediction: *'der sieg war keine grosse sache '*

(4.21) **WER: 22.2% | CER: 1.64% | BLEU: 0.624**
Transcription: *'die restlichen 57 prozent wurden im zivilen bereich generiert'*
Prediction: *'die restlichen 57 prozent wurden im zivilenbereich generiert'*

(4.22) **WER: 140.0% | CER: 78.95% | BLEU: 0.0**
Transcription: *'doch die kritischen stimmen überwiegen '*
Prediction: *'aber die kritischen stimmen sind dan klar in eier mehrheit'*

The examples 4.23 to 4.26 present some predictions of the Whisper model fine-tuned on the larger prepared speech data set. Similarly to XLS-R, the model performs better if fine-tuned on more data. The small version achieves a word error rate of 86.49%, a character error rate of 63%, and a BLEU score of 0.036 if applied to prepared speech data. These metrics are significantly lower if the size of the data set is increased. One main problem with the large version of the fine-tuned Whisper is that it predicts the letter 'c' at the beginning of some sentences, as illustrated in the examples. The example with the worst word error rate shows a prediction where the ground-truth differs massively from the sentence produced by the model.

Large Model Whisper – Trained on Prepared Speech

(4.23) **WER: 33.3% | CER: 13.64% | BLEU: 0.452**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'cdie spannende verfassung erlaubt die abspaltung einer der region nicht'*

(4.24) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'das war ein tolles erlebnis'*
Prediction: *'das war ein tolles erlebnis'*

(4.25) **WER: 16.67% | CER: 2.22% | BLEU: 0.7598**
Transcription: *'auch intern stehen mogliche kandidaten bereit'*
Prediction: *'cauch intern stehen mogliche kandidaten bereit'*

(4.26) **WER: 160.0% | CER: 122.2% | BLEU: 0.0**
Transcription: *'und die tribune bleibt leer'*
Prediction: *'zen in der folge wurden jedoch nicht erhoben'*

4.3 Impact of the Default Parameters

The developers of the Whisper suggested only one gradient accumulation step instead of two, as suggested for XLS-R. The result is a smaller training batch size. Furthermore, the suggested learning rate is 0.00001 instead of 0.0003. Since these parameters have to be changed to make the model comparable to the XLS-R, some other versions of Whisper are fine-tuned to investigate what impact these parameters have on the performance.

Table 4.4 shows the performance of three versions with the changes in the parameters mentioned above. The saving step size is set to 100 and is kept the same for all three versions because, for training Whisper on the large version of the prepared speech data set, any larger saving step size would take a long time to fine-tune and force the system to crash. Therefore, the size of the saving step is set to 100 for all three versions. One version is trained on the large prepared speech data set version, containing 84'771 samples. The second version was trained on the small subset of the prepared speech data set to make it comparable to the initial, comparable version of the Whisper. The third version is fine-tuned on the spontaneous speech data set that contains 1857 samples.

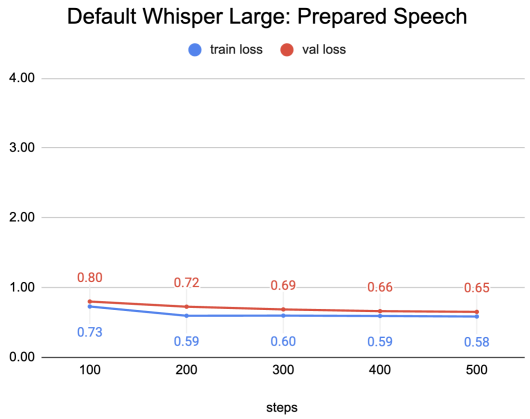
Default Whisper Large		**Default Whisper Small**		**Default Whisper**	
Prepared Speech		**Prepared Speech**		**Spontaneous speech**	
WER	42.57%	**WER**	38.66%	**WER**	41.51%
CER	24.12%	**CER**	21.9%	**CER**	12.07%
BLEU	0.412	**BLEU**	0.44	**BLEU**	0.3497

Table 4.4: The impact of the step size, training batch size and learning rate on the Whisper model on three data sets.

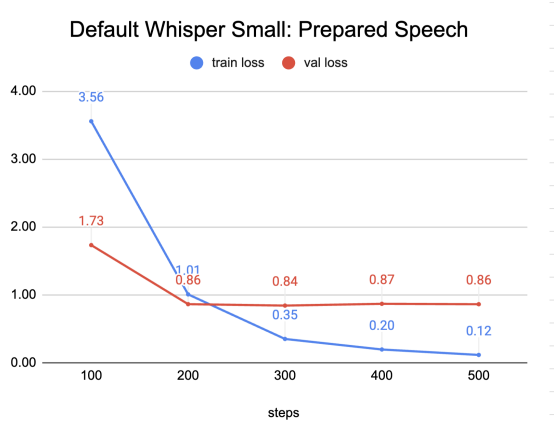
The comparable Whisper models with a training batch size of 32 and a learning rate of 0.0003 have a word error rate of 91.89% if they are trained and tested on the prepared speech data set and a word error rate of 67.02% if they are trained on the spontaneous speech data set. The character error rate is 64.64% for the prepared speech model and 33% for the spontaneous speech model. The error rates of the versions of all data sets drop as the model gets fine-tuned using a setting closer to the default. Surprisingly, the model fine-tuned on the small subset of the prepared speech data outperforms the other two versions.

Figure 4.4 present the loss and error rates of the default Whisper models during the training process. The word error rate of the small version of prepared speech reports a massively increasing word error rate. It starts at 72% and increases to 1230% over 500 steps. Nevertheless, this model reports the best word error rate and

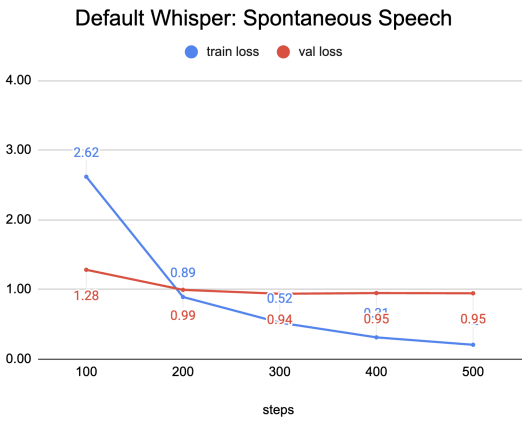
BLEU score compared to the other two default Whisper models. This can be due to an error in the system in calculating the word error losses during the training process. In Figure 4.4d, these outliers are only partially reported and marked as a yellow dotted line.



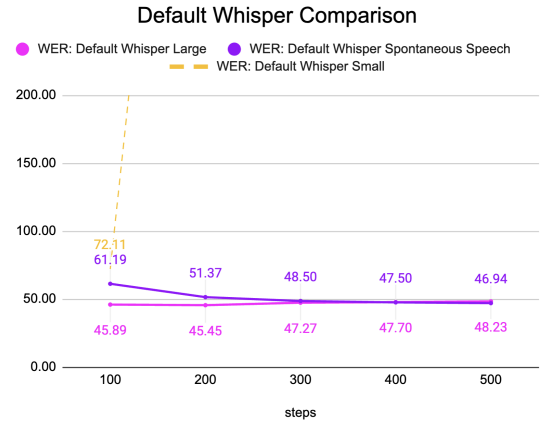
(a) Train and validation loss of the default Whisper, fine-tuned on the large prepared speech data set.



(b) Train and validation loss of the default Whisper, fine-tuned on the smaller prepared speech data set.



(c) Train and validation loss of the default Whisper, fine-tuned on spontaneous speech data.



(d) Word error rate the Whisper models trained with default parameters. The unit of the word error rate is displayed in percentage.

Figure 4.4: Training report of the Whisper models trained with default parameters.

The examples 4.27 to 4.30 illustrate some sample predictions of the default Whisper model that was fine-tuned on the large prepared speech data set. There are many predictions in the test set with an error rate close to zero. If the word error rate is higher, one main reason is dividing a word into two parts by inserting a white space. The prediction 4.30 is only one example where the model transcribes the sentence in English. This phenomenon will be discussed in Chapter 5.

Large Default Model Whisper – Trained on Prepared Speech

(4.27) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*

(4.28) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'seit dem konzil gehen wir noch starker in diese richtung'*
Prediction: *'seit dem konzil gehen wir noch starker in diese richtung'*

(4.29) **WER: 28.57% | CER: 2.44% | BLEU: 0.680**
Transcription: *'ist eure basis rund die uhr einsatzbereit'*
Prediction: *'ist eure basis rund die uhr einsatz bereit'*

(4.30) **WER: 254.545% | CER: 130.99% | BLEU: 0.0**
Transcription: *'noch sind die daten uber das gesamte jahr nicht vollstandig ausgewertet'*
Prediction: *'at the moment the data for this whole year is not yet fully at the moment the data for the whole year is not yet fully worth it'*

The examples 4.31 to 4.34 show some predictions of the small default Whisper model trained on prepared speech. Prediction 4.34 is an example of the model repeating the transcribed text. This seems to be an individual case for this model as it does not tend to hallucinate or repeat words if the whole output on the test set is analysed.

Small Default Model Whisper – Trained on Prepared Speech

(4.31) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*

(4.32) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'fur uns ist der text der verordnung entscheidend'*
Prediction: *'fur uns ist der text der verordnung entscheidend'*

(4.33) **WER: 28.57% | CER: 2.22% | BLEU: 0.455**
Transcription: *'eine leser reporterin hat den einsatz gefilmt'*
Prediction: *'eine leserreporterin hat den einsatz gefilmt'*

(4.34) **WER: 2554.55% | CER: 2097.1% | BLEU: 0.0**

Transcription: *'danemark ist einem uno bericht zufolge das glucklichste land der welt'*

Prediction: *'denn die marke ist gemessen und man nun bericht denn die marke ist gemessen und man nun bericht denn die marke ist gemessen und man nun bericht [...] denn die marke ist gemessen und mann denn die marke'*

The default Whisper model is also fine-tuned for spontaneous speech. This version has similar problems to models presented earlier in this chapter, as it struggles with recognising the beginning and the end of a word.

Default Model Whisper – Trained on Spontaneous Speech

(4.35) **WER: 28.57% | CER: 5.71% | BLEU: 0.3499**

Transcription: *'hat er min reschpakt wil daas isch gaar kas chliises amtli das isch es'*

Prediction: *'hat er miin reschpakt will daas isch gaarka s chliises amtli das isch es'*

(4.36) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**

Transcription: *'beschaftigt au im zamehang mit dem was di letschte zwai wuche passiert isch ich bin uberzuugt'*

Prediction: *'beschaftigt au im zamehang mit dem was di letschte zwai wuche passiert isch ich bin uberzuugt'*

(4.37) **WER: 27.27% | CER: 3.39% | BLEU: 0.0**

Transcription: *'das ebe schurnalischte s gfuul hand mer chong aim ales sage'*

Prediction: *'das ebe schurnalischtes gfuul hand mer chond aim ales sage'*

(4.38) **WER: 133.3% | CER: 24.0% | BLEU: 0.0**

Transcription: *'zwaihundertsachzgmilioone schwiizerfranke choschte'*

Prediction: *'zwai hundertsachzg milioone schwiizefranken'*

4.4 Improvement Strategies on Spontaneous Speech

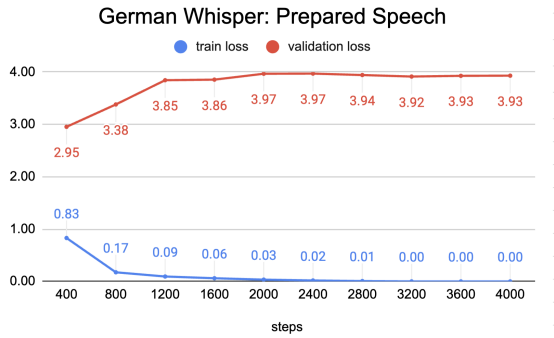
This chapter presents the results of some improvement strategies applied to the Whisper model fine-tuned on spontaneous speech. Each model is trained independently of the other models.

4.4.1 Specify the Language

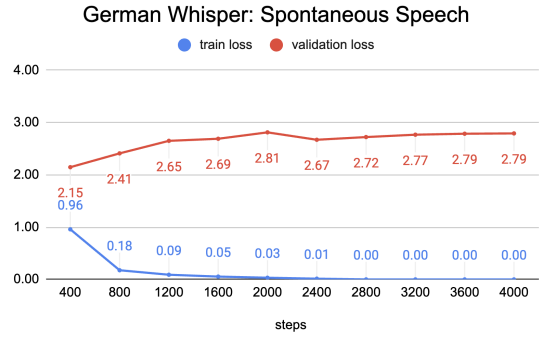
The first strategy is the specification of the language. The Whisper tokeniser can take a language token to specify the transcription language. As the model tends to predict English sentences in some cases, one hypothesis is that adding the language token leads to better performance. In the case of spontaneous speech, the performance drops if the German language token is added to the tokeniser. Important to mention is that the transcription of spontaneous speech is in Swiss German while the transcription of the prepared speech data is Standard German. To make the results comparable, the small subset of the prepared speech data is used, and the preprocessing and the parameters are kept the same as for the comparable Whisper version. The results of Section 4.1 are added to Table 4.5 to visualise the performance. As the table shows, for the prepared speech, the model performs very similarly, independent of the language token. For spontaneous speech, the language token even leads to worse performance.

Comparable Whisper without German language token				**Comparable Whisper with German language token**			
Prepared Speech		**Spont. Speech**		**Prepared Speech**		**Spont. Speech**	
WER	86.49%	**WER**	67.02%	**WER**	86.80%	**WER**	99.47%
CER	63.80%	**CER**	33%	**CER**	62.37%	**CER**	81.10%
BLEU	0.036	**BLEU**	0.156	**BLEU**	0.058	**BLEU**	0

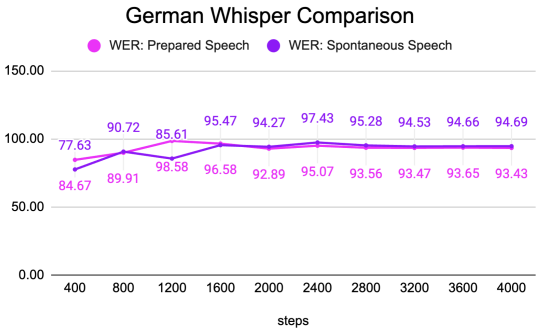
Table 4.5: The impact of adding the German language token to the prepared or spontaneous speech Whisper models. The results are reported for the test set of the corresponding speech style. For comparison, the results of the comparable Whisper versions without the German language token are added to the table.



(a) Train and validation loss of the German Whisper, fine-tuned on the small prepared speech data set.



(b) Train and validation loss of the German Whisper, fine-tuned on the spontaneous speech data set.



(c) Word error rates of the Whisper models trained with the tokeniser containing a German language token. One model trained on prepared speech and the other model trained on spontaneous speech. The unit of the word error rate is displayed in percentage.

Figure 4.5: Training report of the Whisper models trained with the German language token.

The examples 4.39 to 4.42 show some good and bad results of the prepared speech Whisper model that is trained with the German language token. In the example with the highest error rates, the model starts hallucinating and repeats the word 'der' multiple times. This is not a single case, as there are several similar predictions for other sentences.

Whisper: German Language Token – Trained on Prepared Speech

(4.39) **WER: 88.89% | CER: 59.09% | BLEU: 0.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'die ist panisch die verfannungs die clubspieler von neuerung'*

(4.40) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'und das ist auch heute noch so'*
Prediction: *'und das ist auch heute noch so'*

(4.41) **WER: 28.57% | CER: 8.696% | BLEU: 0.615**
Transcription: *'denn nur das ist wirklich politische literatur'*
Prediction: *'ratdenn nur das ist wirklich politische literature'*

(4.42) **WER: 2377.78% | CER: 1097.37% | BLEU: 0.0**
Transcription: *'doch seine administration scheiterte am vereinten widerstand der schmarotzer'*
Prediction: *'leichdie minder der der der der der der der der der der der der der der der der der der der der der [...] der der der der der der'*

The model trained on prepared speech is hallucinating, as well as the model trained on spontaneous speech. Examples of the model output in the test set demonstrate the low performance visible in Table 4.5. Again, the model is hallucinating and repeating some words or tokens. If the model is not hallucinating, the prediction is significantly different from the ground-truth, resulting in not only a high word error but also a high character error rate. For this reason, no example is provided showing a high word error rate but a low character rate, as there are no such examples in the test set.

Whisper: German Language Token – Trained on Spontaneous Speech

(4.43) **WER: 100.0% | CER: 77.14% | BLEU: 0.0**
Transcription: *'hat er min reschpakt wil daas isch gaar kas chliises amtli das isch es'*
Prediction: *'irgendwie uf dorfiifef ghort hat er'*

(4.44) **WER: 75.0% | CER: 61.70% | BLEU: 0.0**
Transcription: *'dass aagriff uf bolizischte a wurklich eso seer'*
Prediction: *'irgendwie uf dam a wuklich ebevoor'*

(4.45) **WER: 500.0% | CER: 440.0% | BLEU: 0.0**
Transcription: *'roschee ich mus inz wukli sage'*
Prediction: *'pt a psch psch psch [...] psch psch psch psch psch psch'*

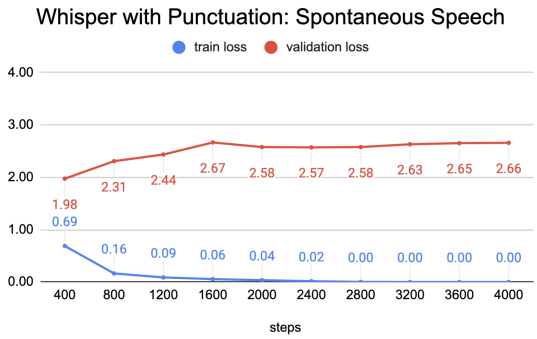
4.4.2 Consider Punctuation for Whisper

Whisper is capable of predicting punctuation. On the contrary, XLS-R does not predict any punctuation. For this reason, for the comparable Whisper version, the special characters are deleted from the training data during the preprocessing before the model was fine-tuned. Therefore, the goal is to test in this improvement step whether keeping the punctuation in the test set leads to better results. Table 4.6 shows similar results for the evaluation metrics as the comparable Whisper model.

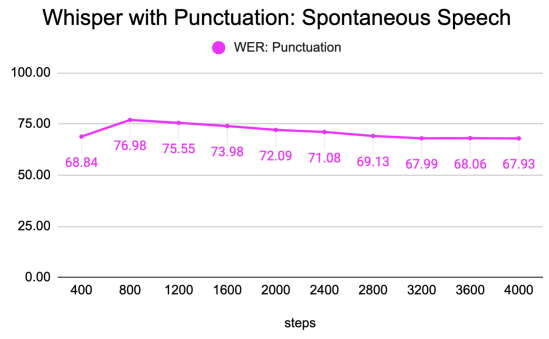
Whisper with Punctuation	
Spontaneous Speech	
WER	68.15%
CER	33.18%
BLEU	0.16%

Table 4.6: The impact of keeping the punctuation in the training set for fine-tuning the Whisper model with spontaneous speech data.

In Figure 4.6, the training process is visualised. Similar to some other versions, the training loss converges to 0 after around 2000 steps, while the validation loss flattens after increasing. The word error rate first increases before it drops by a few percentage points.



(a) Train and validation loss of the Whisper model fine-tuned on the spontaneous speech data set, keeping the punctuation during the preprocessing.



(b) Word error rate of the Whisper model fine-tuned on the spontaneous speech data set, keeping the punctuation during the preprocessing, in percentage.

Figure 4.6: Training report of the Whisper models trained while keeping the punctuation during the preprocessing.

The examples 4.46 to 4.49 show some predictions of the Whisper model that keeps the punctuation. As visible in the examples 4.46 and 4.49, as well as in other predictions of the test data, the model often starts sentences with "els". After this starting sequence of characters, the sentence often continues correctly or with few errors. This leads to a high word error rate but a lower character error rate.

(4.46) **WER: 78.57% | CER: 35.71% | BLEU: 0.0**
Transcription: *'hat er min reschpakt wil daas isch gaar kas chliises amtli das isch es'*
Prediction: *'elshat berliine reschpaktli daas isch d bar ka grosgeli eso entli das isch es'*

(4.47) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'und de phunkt isch afach daa'*
Prediction: *'und de phunkt isch afach daa'*

(4.48) **WER: 66.67% | CER: 7.69% | BLEU: 0.0**
Transcription: *'das agit sich'*
Prediction: *'das a git sich'*

(4.49) **WER: 200.0% | CER: 42.0% | BLEU: 0.0**
Transcription: *'zwaihundertsachzgmilioone schwiizerfranke choschte'*
Prediction: *'elszwaihundertzachzgi role schwiizer fall chommer choo'*

4.4.3 Disfluency Labelling

This section presents the results of the Whisper model if the special annotations are not removed during the preprocessing. These special annotations are also called disfluency labels, as they mark hesitations, noise, and other sounds that are not speech. This strategy was tested only in spontaneous speech because the special annotations are only part of this data set. The prepared speech does not have information about disfluency, as this is not common in prepared speech. For the first version, keeping the special annotations is the only difference from the comparable spontaneous speech Whisper model. As some curves of the loss seem to be overfitting for the models trained on spontaneous speech, dropout is added to a second version. As Table 4.7 shows, adding a dropout of 0.1 significantly increases performance. Dropout is not the only thing that is added to the model to avoid overfitting. In addition, the language token is added not only to the tokeniser but also to the processor. Moreover, a separate model configuration forces the model to predict in German:

model.generate = partial(model.generate, language='german', task='transcribe')

Whisper with Disfluency labelling		Whisper with Disfluency Labelling and Dropout	
Spontaneous Speech		**Spontaneous Speech**	
WER	83.54%	**WER**	68.65%
CER	47.79%	**CER**	35.35%
BLEU	0.119	**BLEU**	0.185

Table 4.7: Impact of adding disfluency labels to the training data of the Whisper model trained and tested on spontaneous speech.

Both plots on the training and validation loss behave the same as the numbers are identical over the 4000 training steps. The word error rate changes differently if both models are compared. For the model that contains only disfluency labels, the word error rate drops over around 1200 training steps, increases again, and levels off during the rest of the training steps. If dropout is added, the word error rate increases for the first 800 steps but does not change afterwards.

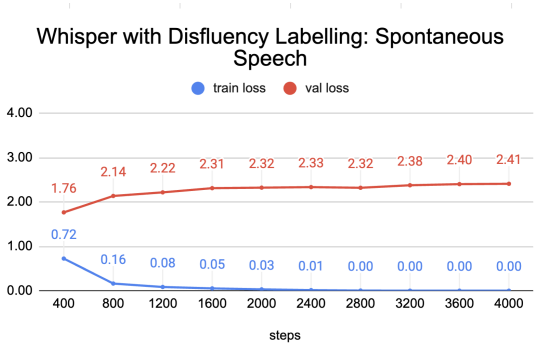
The model trained to recognise disfluency labels often predicts the letters 'od' at the beginning of a sentence. The worst result shows another example where the model repeats a word multiple times. It is recognisable that the model is capable of predicting the disfluency label 'breath-mouth-noise'. The other special annotations containing non-alphanumerical characters are not visible in the output due to the normalisation step described in Chapter 3.4.

Whisper with Disfluency Labelling – Trained on Spontaneous Speech

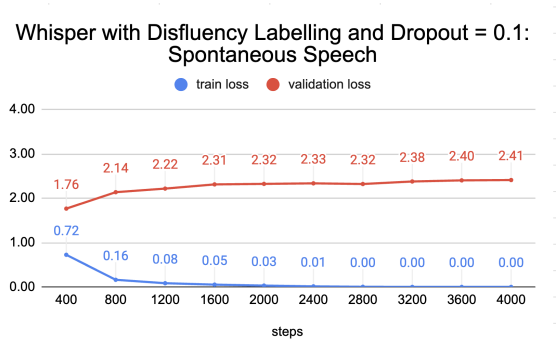
(4.50) **WER: 80.0% | CER: 41.38% | BLEU: 0.0**
Transcription: *'hat er min reschpakt wil daas breathmouthnoise isch gaar kas chliises amtli das isch es'*
Prediction: *'odhat kha s chum ireschpruch gul daas breathmouthnoise isch d park hand s so ires das isch es'*

(4.51) **WER: 11.1% | CER: 7.89% | BLEU: 0.88**
Transcription: *'au mit de koleege zame wo mer ois dann'*
Prediction: *'od au mit de koleege zame wo mer ois dann'*

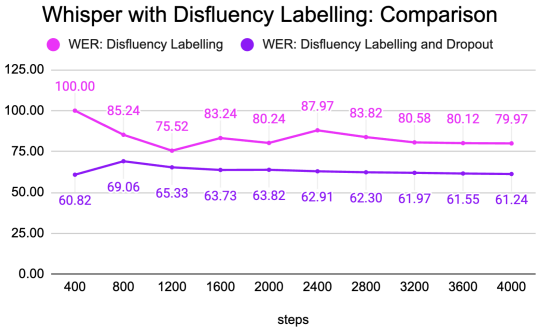
(4.52) **WER: 20.0% | CER: 4.0% | BLEU: 0.7598**
Transcription: *'schriftlich beschtaatiged han ich i de eerschte paar wuche breathmouthnoise'*
Prediction: *'odschriftlich beschtaatiget han ich i de eerschte paar wuche breathmouthnoise'*



(a) Train and validation loss of the Whisper, fine-tuned on the small spontaneous speech data set containing disfluency labels without dropout.



(b) Train and validation loss of the German Whisper, fine-tuned on the spontaneous speech data set containing disfluency labels and a dropout of 0.1.



(c) Word error rates of the Whisper models trained using the disfluency labels on spontaneous. One model is implemented with dropout. The word error rates are reported in percentage.

Figure 4.7: Training report of the Whisper models trained on spontaneous speech containing disfluency labels.

(4.53) **WER: 2170.0% | CER: 1783.6% | BLEU: 0.0**

Transcription: *'seer aa zrugghaltend daa ggurtailt wird und alafalls gschp aa'*

Prediction: *'odbreathmouthnoise seer aa druk hals daa gtait wird ailt viert viert wurt wurt wurt wurt wurt wurt wurt wurt [..] wurt wurt wurt wurt wurt wurt wurt wurt wurt wurt wurt'*

The examples 4.54 to 4.57 illustrate some examples of the predictions given by the Whisper model that is trained on the disfluency labels and a dropout of 0.1. Again, the model tends to repeat some words. In the case of Example 4.57, the first part of the prediction is correct. The model begins to struggle and repeats itself as the

words in the ground-truth become very similar. There are no recognisable characters that the model adds to the sentences as the other Whisper disfluency model does by adding 'od'.

Whisper with Disfluency Labelling, Dropout and Generator for Whisper – Trained on Spontaneous Speech

(4.54) **WER: 40.0% | CER: 11.49% | BLEU: 0.0**
Transcription: *'hat er min reschpakt wil daas breathmouthnoise isch gaar kas chliises amtli das isch es'*
Prediction: *'hat er minere schpaggs wi daas breathmouthnoise isch d ggar kas chliises antli das isch es'*

(4.55) **WER: 0.0% | CER: 0.0% | BLEU: 1.0**
Transcription: *'mir chond ois afacht nume erlaube breathmouthnoise'*
Prediction: *'mir chond ois afacht nume erlaube breathmouthnoise'*

(4.56) **WER: 28.57% | CER: 2.17% | BLEU: 0.54**
Transcription: *'ich han breathmouthnoise so vill presidie scho'*
Prediction: *'ich han breathmouthnoise so vill presi die scho'*

(4.57) **WER: 1661.54% | CER: 1165.75% | BLEU: 0.018**
Transcription: *'ooni name z nane breathmouthnoise das wann opper sich noch nod nod mochti'*
Prediction: *'ooni nacher zuene breathmouthnoise das wann opper sich nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod nod [...] nod nod nod nod nod nod n'*

For both models, there are some samples in the test set where the transcription is labelled with the tag [no-relevant-speech]. In this case, the model does not predict this tag, but it tries to transcribe the noisy audio. The example below illustrates this case.

(4.58) **WER: 2600.0% | CER: 677.78% | BLEU: 0.0**
Transcription: *'[no-relevant-speech]'*
Prediction: *'music ja ich hund dann daa de mues nod noo vor ine wo wo wo wo si bisch dann sfaarnseerschted breathmouthnoise da kli ir nuut kaarned'*

4.5 Zero-Shot Learning

The Whisper models developed by OpenAI are pre-trained and allow zero-shot learning where the model is used without fine-tuning. Therefore, the performance of the models is measured when the model is used on the data set without being fine-tuned on the prepared and spontaneous speech data. To compare XLS-R and Whisper, the zero-shot performance is tested on both pre-trained models. From tables 4.8 and 4.9, it can be noted that XLS-R performs worse than Whisper when it comes to zero-shot usage of the model. Based on these results, possible improvement strategies are tested by taking the Whisper model as a base.

4.5.1 Wav2Vec2-XLS-R Zero-Shot

Table 4.8 shows that for the XLS-R model, the word error rate does not exceed 100% if the model is used without fine-tuning. The character error rate is massively higher for the XLS-R zero-shot evaluation than for the Whisper model used under the same conditions.

XLS-R Zero Shot			
Prepared Speech		**Spontaneous Speech**	
WER	100%	**WER**	100%
CER	188.66%	**CER**	200.89%
BLEU	0	**BLEU**	0

Table 4.8: Test results when XLS-R is used without being fine-tuned on the data sets in percentage.

Example 4.59 presents the reason for the poor performance. The model predicts only some random characters. The other sample outputs on the test set look similar. The only difference is the order of the sequence of characters and the length of the prediction. The length of the transcription and the length of the prediction seem to have no correlation.

XLS-R Zero-Shot Prepared Speech

(4.59) **WER: 100% | CER: 146.97% | BLEU: 0.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'icicjc[UNK]j[UNK]cp[UNK]cwcjcdcjcjc[UNK]cwc[UNK]cw[UNK]d[UNK]cw[UNK]c[UNK]cpcwcd[UNK]c[UNK]cwcwcdcjccch'*

The evaluation of the XLS-R model using zero-shot learning in spontaneous speech shows results similar to zero-shot learning in prepared speech. The model predicts some random characters where the length of the generated prediction does not correlate with the length of the transcription, as Example 4.60 shows.

XLS-R Zero-Shot Spontaneous Speech

(4.60) **WER: 100% | CER: 253.3% | BLEU: 0.0**
Transcription: *'jaa also bis jez'*
Prediction: *'aã[UNK]mã[UNK]ãh[UNK]hãmtgaã[UNK]hãzaht'*

4.5.2 Whisper Zero-Shot

Whisper is also tested for its zero-shot performance. As presented in Table 4.9, the results are similar for spontaneous and prepared speech. But both versions outperform the XLS-R zero-shot, as the character error rate is clearly lower for the Whisper zero-shot learning than for the XLS-R zero-shot learning. The word error rate is not comparable since, for XLS-R, the maximum word error rate cannot exceed 100%.

Whisper Zero Shot			
Prepared Speech		**Spontaneous Speech**	
WER	110.74%	**WER**	105.33%
CER	70.25%	**CER**	73.99%
BLEU	0.0064	**BLEU**	0

Table 4.9: Test results when Whisper is used without being fine-tuned on the data sets in percentage.

The first example is the reference prediction of Example 4.59 to see the difference between the outputs of the models on the same input audio. Example 4.62 is one of the samples with the lowest word and character error rates and the best BLEU score. In Example 4.63, the character error rate is lower than the word error rate,

but only one word is incorrect because it is transcribed in English. The last example shows the prediction with the highest word and character error rate. The error rates are that high because the model starts hallucinating in the middle of the sentence and lines up the sequence 'land based'

Whisper Zero-Shot Prepared Speech

(4.61) **WER: 100% | CER: 63.64% | BLEU: 0.0**
Transcription: *'die spanische verfassung erlaubt die abspaltung einer region nicht'*
Prediction: *'the spanish attachment allows the maintenance of a region'*

(4.62) **WER: 0% | CER: 0% | BLEU: 1.0**
Transcription: *'sie werden alle der spionage beschuldigt'*
Prediction: *'sie werden alle der spionage beschuldigt'*

(4.63) **WER: 20% | CER: 12.5% | BLEU: 0.67**
Transcription: *'der druck aus bern wirkt'*
Prediction: *'the druck aus bern wirkt'*

(4.64) **WER: 2950.0% | CER: 2228.17% | BLEU: 0.0**
Transcription: *'die selbstmord rate von landwirten liegt deutlich uber dem durchschnitt'*
Prediction: *'the self moderate of land based land based land based land based land based land based land based land based land based [...] land based land'*

The examples 4.65 to 4.68 show the performance of zero-shot learning on spontaneous speech data. The first example is a prediction where the English and German languages are mixed. The second one predicts a Standard German sentence, which results in a lower character error rate, but a higher word error rate due to the given Swiss German transcription. The third example shows an English prediction with an extraordinarily high word and character error rate. Regarding the sentence's content, the prediction and transcription are close.

Whisper Zero-Shot Spontaneous Speech

(4.65) **WER: 75% | CER: 40% | BLEU: 0.0**
Transcription: *'jaa also bis jez'*
Prediction: *'yeah also bis jetzt'*

(4.66) **WER: 66.67% | CER: 46.67% | BLEU: 0.0**
Transcription: *'in afrikaa sind'*
Prediction: *'in africa'*

(4.67) **WER: 87.5% | CER: 22.81% | BLEU: 0.0**
Transcription: *'empfole mer mussti en offensiivschtrategii faare und sage'*
Prediction: *'empfohlen man musste ein offensivstrategie fahren und sagen'*

(4.68) **WER: 200% | CER: 111.76% | BLEU: 0.0**
Transcription: *'ganz eso dramatisch isch das wider'*
Prediction: *'the whole thing is so dramatic that it is all over again'*

5 Discussion

Chapter 5 first compares the results of Chapter 4. The final parts of the chapter discuss the limitations of the thesis.

5.1 Interpretation of the Results

The following section will summarise the results presented in the last chapter and gives an overview of all the models that are developed to compare prepared and spontaneous speech.

5.1.1 Comparison of the Models

Prepared Speech			
Ground-Truth	die spanische verfassung erlaubt die abspaltung einer region nicht		
Evaluation on the Whole Test Set	**Model**	**Reference Prediction**	**Evaluation of the Reference Prediction**
WER: 79. 95 CER: 34.48 BLEU: 0.054	**Comparable XLS-R**	die spanischeven fassung erlaubt die abspaltung einn den ereregion nicht	WER: 55.5 CER: 15.15 BLEU: 0.0
WER: 36.03 CER: 16.46 BLEU: 0.422	**Large XLS-R**	die spanische verfassung erlaubt die abspaltung eine einer region nicht	WER: 11.11 CER: 7.576 BLEU: 0.658
WER: 86.49 CER: 63.799 BLEU: 0.063	**Comparable Whisper**	st spanische verf in der lebt poststen viertig	WER: 88.89 CER: 56.06 BLEU: 0.0
WER: 54.45 CER: 45.19 BLEU: 0.233	**Large Whisper**	cdie spannende verfassung erlaubt die abspaltung einer der region nicht	WER: 33.3 CER: 13.64 BLEU: 0.452
WER: 42.57 CER: 24.12 BLEU: 0.412	**Default Whisper Large**	die spanische verfassung erlaubt die abspaltung einer region nicht	WER: 0.0 CER: 0.0 BLEU: 1.0
WER: 38.66 CER: 21.9 BLEU: 0.44	**Default Whisper Small**	die spanische verfassung erlaubt die abspaltung einer region nicht	WER: 0.0 CER: 0.0 BLEU: 1.0
WER: 86.80 CER: 62.37 BLEU: 0.058	**German Whisper**	die ist panisch die verfannungs die clubspieler von neuerung	WER: 88.89 CER: 59.09 BLEU: 0.0

Table 5.1: Comparison of all models that are trained on prepared speech. The table reports the WER, CER and BLEU scores on the whole test set and the reference prediction. The WER and CER are reported in percentage.

Tables 5.1 and 5.2 show the reference sentences of the prepared and spontaneous speech data on all the models presented in the preceding section. These tables should give an overview of the performance of the final versions of the models that are developed in the scope of this thesis. While the second column marks the name of the model, the first column reports the performance of the corresponding model over the whole test set of the given speech style. The third column reports the prediction of the models on the reference audio. The column on the right side reports the error rates and BLEU score of the specific reference sentence.

Spontaneous Speech			
Ground-Truth XLS-R	hät er min reschpäkt wil daas isch gaar käs chliises ämtli das isch es		
Ground-Truth Whisper	hat er min reschpakt wil daas isch gaar kas chliises amtli das isch es		
Ground-Truth Disfluency Labelling	hat er min reschpakt wil daas breathmouthnoise isch gaar kas chliises amtli das isch es		
Evaluation Metrics on the Whole Test Set	**Model**	**Reference Prediction**	**Evaluation of the Reference Prediction**
WER: 60.89 CER: 16.74 BLEU: 0.178	**Comparable XLS-R**	hät er minreschpäkt will daas isch gaarkäs s chliis es aämtli das isch es	WER: 57.14 CER: 10.0 BLEU: 0.0
WER: 67.02 CER: 33 BLEU: 0.156	**Comparable Whisper**	hat er miere schpakt i daas isch parargkas chliese s andli das isch es	WER: 50.0 CER: 17.14 BLEU: 0.0
WER: 41.51 CER: 12.07 BLEU: 0.3497	**Default Whisper**	hat er miin reschpakt will daas isch gaarka s chliises amtli das isch es	WER: 28.57 CER: 5.71 BLEU: 0.3499
WER: 99.47 CER: 81.105 BLEU: 0.0	**German Whisper**	irgendwie uf dorfiifef ghort hat er	WER: 100.0 CER: 77.14 BLEU: 0.0
WER: 68.154 CER: 33.18 BLEU: 0.16	**Punctuation Whisper**	elshat berliine reschpaktli daas isch d bar ka grosgeli eso entli das isch es	WER: 78.57 CER: 35.71 BLEU: 0.0
WER: 83.54 CER: 47.79 BLEU: 0.119	**Disfluency Whisper**	odhat kha s chum ireschpruch gul daas breathmouthnoise isch d park hand s so ires das isch es	WER: 80.0 CER: 41.38 BLEU: 0.0
WER: 68.65 CER: 35.35 BLEU: 0.185	**Disfluency and Dropout**	hat er minere schpaggs wi daas breathmouthnoise isch d ggar kas chliises antli das isch es	WER: 40.0 CER: 11.49 BLEU: 0.0

Table 5.2: Comparison of all models that are trained on spontaneous speech. The table reports the WER, CER and BLEU scores on the whole test set and the reference prediction. The WER and CER are reported in percentage.

5.1.2 Explanations of the Results

The results show that for the comparable version, the spontaneous speech model performs better than the prepared speech model. This is true for XLS-R and Whisper. This contradicts the research presented in 2, which proved a performance drop for spontaneous speech. One possible reason is the format of the data used. For both speech styles, short texts are used to fine-tune the model. However, especially spontaneous speech should be trained and tested on longer speech sequences for a more natural setting. It can be assumed that testing the spontaneous speech XLS-R and Whisper on longer audio would decrease the performance. In the case of the data

used, the two data sets could be too similar in format. However, the spontaneous speech data consist of hesitations, fillers, and many incomplete sentences. This leads to another assumption that spontaneous speech is not necessarily as scalable as prepared speech. For prepared speech, a large data set was available. Therefore, the impact of using more data to fine-tune the model could be proven. It is not clear whether expanding the spontaneous speech data set would have the same impact.

5.1.3 Problems

Chapter 4 presents many examples of the predicted transcriptions that a model gives when it receives audio as input. Throughout all the examples, some problems are not specific to one particular version but seem to run through all the models that were produced within the scope of this thesis.

Hallucinating and Repetition Looking at the worst prediction of the models, hallucinations or random repetitions of words are one reason that reduces the overall quality of the model evaluated on the test set. These errors increase the word and character error rate and let the BLEU score drop to 0. Some models repeat only one random word that is not part of the audio, which is called hallucination. Other models start to transcribe the audio correctly, but repeat one or more words of the sentence multiple times. In both cases, the predicted sentence becomes massively longer than the ground-truth leading to a remarkably high word error rate of over 1000%. These outliers have a direct impact on the word error rate that was reported on the whole test set. Whisper reports hallucination as a limitation on their model description on Huggingface (OpenAI, 2023). They are reasoning this error with the fact that the models are pre-trained in a weakly supervised manner and because they use large-scale noisy data (OpenAI, 2023). As the developers of Whisper further explain, they assume that models try to predict the next word in the audio while transcribing based on their general knowledge of language (OpenAI, 2023). Moreover, they assume that this problem will increase for low-resource languages (OpenAI, 2023)

Long Vowels in Spontaneous Speech The transcriptions in the spontaneous speech data set are written in a standardised Swiss German based on the Dieth rules (Dieth and Schriftkommission, 1938). According to these rules, long vowels are written double-spelt. For example, the letter ‘a’ in ‘aafange’ (en: start/begin) is duplicated. As some examples in Chapter 4 showed, the model is not always able

to detect long vowels. Therefore, the vowel is part of the transcription, but is not duplicated. This leads to a high word error rate but a lower character error rate.

Handling Digits There exist sentences in the test set where numbers are mentioned by the speaker. Moreover, this is a real scenario if a system has to transcribe speech to text. Depending on the annotator, numbers are written in words, or some digits are added to the text. For the model, it is difficult to assess if it should write a number in words, especially if it is pre-trained on data collected from different domains. Moreover, it is hardly possible to be consistent in a data set that is used to fine-tune and evaluate the system. Differences in formatting numbers lead to lower performance if word or character error rates are used as an evaluation metric.

Defining Word Boundaries Almost all models seem to struggle with word boundaries. Sometimes, they add white space in between a word and split the word apart. There also exist opposite examples where the model concatenates two words into one. Single characters that belong to the word on the other side of the white space are another consequence of this word boundary problem.

Prediction in English Some models tend to predict sentences in English even if the audio is in Swiss German, and they are fine-tuned on a data set containing only Swiss German audio and (Swiss) German transcriptions. Noticeably, in the case of the English transcription, the statement or idea behind the sentence is often correct. There are errors in grammar, as it is sometimes an English word-by-word translation of the audio, but it is recognisable what the speaker was trying to say. Examples 5.1 and 5.2 are exemplary predictions by the large default Whisper model trained on prepared speech. This model is not the only one that produces such results, as there exist other versions containing similar predictions on the test set.

(5.1) **WER: 237.5% | CER: 130.36% | BLEU: 0.0**
Transcription: *'ebenfalls seit zwei jahren rucklaufig sind die einnahmen'*
Prediction: *'eitherfalls in 2 years time are they even though they are 2 years worth of return they are enormous'*

(5.2) **WER: 155.56% | CER: 97.96% | BLEU: 0.0**
Transcription: *'damit liegt die zahl funfmal hoher als im vorjahr'*
Prediction: *'in this case the number is 5 times higher than in the previous year'*

5.2 Obstacles

The following section describes the obstacles and challenges that had to be overcome during the work.

5.2.1 Resources and Data Processing

One obstacle is the time required for the models to preprocess and train. To use GPU and TPU and for a better code structure, Google Colab was used to develop and run the code. However, the preparation of the data for training initially exceeded the memory and storage capacity. Moreover, converting the audio to a machine-readable format was too time consuming. The loading of the data with Numpy[1] arrays or Pandas[2] data frames did not accelerate the process. Since inputting the data to the models for training is easier if the data is saved in a Huggingface data set, the solution was to load the prepared and spontaneous speech data sets directly to a repository on Huggingface Hub. Doing this step in Google Colab was too time consuming, as each data file was looked up separately during the creation of the Huggingface dataset. Moving this process to the local machine seemed to be the fastest solution. Afterwards, the data could be loaded directly from Huggingface into Google Colab for further processing and training. Since Huggingface datasets are optimised for efficient data retrieval, the memory and storage issues were resolved. Preparing the audio for Whisper used more disk space; thus, only audio with a duration below five seconds could be used for Whisper to reduce the size of the data set. Details about Huggingface data sets and data set sizes after preprocessing are described in Section 3.

[1] `https://numpy.org/`

[2] `https://pandas.pydata.org/`

5.2.2 Applicability

During the process, the models were tested not only on their corresponding test set but also in the opposite test set. This means that the prepared speech model was additionally evaluated if the input is spontaneous speech and vice versa. This information is not as informative as the other results because the audio transcripts contained by the data sets are not in the same language. The prepared speech transcriptions are Standard German, while the spontaneous speech transcriptions are an adapted version of Swiss German. The results of these evaluations can be found in the Appendix in Section B.5.

5.3 Limitations

This section describes the limitations of this thesis. This is not just about future work and open questions, as these topics will be discussed in Chapter 6. The goal of this section is to summarise some general limitations that defined the extent of the experiments in Chapter 4.

5.3.1 Computational Resources

In chapter 3.1, the research of Plüss et al. (2022) and Schraner et al. (2022) was mentioned. They used the XLS-R models as well, with the difference in the availability of better computational resources and the time to train larger models. In the scope of this thesis, training huge models with a great number of parameters was not possible due to time and computational power. In the mentioned research, the models were trained on multiple data sets, where not all data was available at the start of the thesis. The first goal was to reproduce the results of Schraner et al. (2022). Due to constraints and differences in time and computational power, this step was discarded because an accurate reproduction of the results was not possible. The resulting step was the development of a different baseline that used a similar architecture and data set. Therefore, the performance of the models is lower compared to recent research. However, for future work, the parameters could be extended, and bigger models with more computational power could lead to better results.

5.3.2 Data

Another limitation is the data and the availability of the data in general. Even with five million Swiss German speakers, it is difficult to collect data from all demographic groups of Swiss German speakers. Additionally, all dialects must be equally distributed in a data set. A few dialects are spoken only in small cantons by a few people. This complicates data collection. The problem of the availability of Swiss German training data was also described by Schraner et al. (2022) as they list publicly available data that can be used to train ASR systems.

Another limitation in spontaneous speech, in general, is the quality of the training data. Training data for spontaneous speech can only be produced by transcribing interviews, television shows, and other conversations (Gabler et al., 2023). The fact that transcription is needed to produce training data for a supervised ASR model leads to noisy data (Gabler et al., 2023). Furthermore, it is difficult to reconstruct spontaneous speech if there are multiple speakers who overlap their sentences (Gabler et al., 2023).

The availability and quality of data lead to another limitation related to the metric of how the performance of a model is evaluated. Transcribing audio is ambiguous, especially in Swiss German, where grammar is not clearly defined, and the word order is not directly applicable to Standard German. Despite there exist multiple ways to transcribe audio, the transcripts can have the same meaning and express the statement made by the speaker. Especially in the case of spontaneous speech, where it is common that sentences are not necessarily complete, and repetitions or fillers are common. If multiple people transcribe audio independently of each other, this could lead to multiple transcriptions. Therefore, it is not clear how a model will behave.

5.3.3 Evaluation Metrics

This thesis uses the word and character error rates, as well as the BLEU score, to evaluate the quality of a model. However, the problems described above show that error rates are not always appropriate for making a statement about the quality of a speech-to-text system. They measure performance quantitatively, comprehensibly, and provably. However, the error rate does not provide any information on the types of errors or the severity of the errors (Morris, 2002). In the previous section, the ambiguity of transcriptions is described, which is one example in which the significance of the given evaluation metrics is limited.

Example 5.3 is a prediction by the large version of the XLS-R model, which is trained on prepared speech.

(5.3) **WER: 140.0% | CER: 78.95% | BLEU: 0.0**
Transcription: *'doch die kritischen stimmen überwiegen'*
Prediction: *'aber die kritischen stimmen sind dan klar in eier mehrheit'*

In this case, the transcription of the model is correct compared to the audio, as the speaker articulates the words predicted by the model. However, ground-truth differs from the audio, even if the statement is correct. As a result, the example is the worst prediction of the model in the test set.

Morris (2002) mentioned two disadvantages of the error rates. On the one hand, the error rate can exceed 100%. Therefore, the models are comparable to each other, but the error rates do not indicate how well a system performs (Morris, 2002). Furthermore, the number of deletions and insertions is not symmetric because, for high error rates, insertions have more weight than deletions (Morris, 2002). The reason can also be seen in Chapter 4 in all examples where the model is hallucinating and the model made many insertions. Furthermore, Huggingface mentions that CER depends on the content and quality of the data set and, therefore, can differ if the same model is used on different data sets (Huggingface, 2023a).

6 Conclusion

The final chapter summarises the findings in the scope of this thesis and discusses where the initial hypothesis and goals are met or refuted. The last section proposes where future work could tie in.

6.1 Summarising the Insights

Given the various versions of the two models and the different improvement strategies, some conclusions can be drawn. One main output of this thesis is the comparison of the performance difference of pre-trained speech models, fine-tuned either on spontaneous or prepared speech. The biggest challenge in achieving this goal was to make the results of the models comparable. On the one hand, the speech styles are contrary, having either Standard German or Swiss German transcription, and differing in their characteristics in general. On the other hand, the architecture and default settings of the hyperparameters of the used state-of-the-art models XLS-R and Whisper differ, resulting in finding an experimental setting to make them comparable as well.

Regarding the two state-of-the-art models, XLS-R and Whisper, the experiments on both models showed that Whisper performs better on zero-shot learning and can be fine-tuned using fewer number of fine-tuning steps. However, training a Whisper model is time consuming. Moreover, XLS-R outperforms Whisper if it is fine-tuned for more steps and, therefore, reported better results as seen in Chapter 4.

It can be said that the size of the training data set has a great impact on the performance of a speech-to-text model. Increasing the size of the data set will reduce the word and character error rate and increase the BLEU score. Moreover, a model is more likely to overfit if the data set is smaller, as the small spontaneous speech data set has proven.

Most of the problems that appear in the setting of spontaneous speech can be found in the output of prepared speech models as well. This leads to the conclusion that

spontaneous speech is not as different from prepared speech as initially assumed. If there are techniques for addressing problems, such as finding word boundaries or handling hallucinations on prepared speech models, these techniques tend to be applicable to spontaneous speech. The experiments do not report problems that are specific to spontaneous speech, but only problems that are specific to the spontaneous speech data set that was used in the scope of this thesis.

It was not possible to show whether a prepared speech model is applicable to spontaneous speech or vice versa, as the two data sets were not of the same format. However, the performance of models fine-tuned on spontaneous speech was significantly higher than for the models that were trained on prepared speech. This insight is promising for future work.

6.2 Future Work

This section describes some points where future work could continue. The scope of the thesis was restricted to a limited period of time. Therefore, there are many open questions and tasks to explore. This chapter should summarise things that were out of the scope of the thesis.

6.2.1 Extension of Resources

The problem with the resources was already described in Chapter 5.2.1. In the thesis, there were fewer resources available, also due to time constraints. Therefore, extending the resources would be the next step. XLS-R and Whisper exist in larger versions that were pre-trained with more parameters. These larger versions could be used to fine-tune. Moreover, the training process could be extended and especially the models where the training and validation loss were still dropping after 4000 steps could be trained longer to achieve better results or to explore at which point the loss starts increasing.

6.2.2 Improvement of the Data

Meanwhile, STT4SG-350, a new data set, was proposed by Plüss et al. (2023). The data set is larger than the data set by Plüss et al. (2022) as it contains 343 hours of speech from all regions of dialect, evenly distributed by gender and age. According to Plüss et al. (2023), it is the largest public speech corpus for Swiss German to

date. As the data set and its paper were just available by July 2023, it was not possible to add the data set to the fine-tuning of the models. For future work, the data set could be combined with the SDS-200 data set and used for fine-tuning. The STT4SG-350 data set also consists of prepared speech data, as the data were collected by showing sentences to speakers and letting them record the sentence in their dialect.

This leads to another point for future work regarding the data. There exist only a few Swiss German spontaneous speech data. One main reason seems to be that transcribing TV shows or discussions is time consuming and costly. In the scope of the thesis, the Schawinski data set was used. The main problem there was that the data set was small compared to the prepared speech data set, and the transcriptions of the two data sets were different. To make the models more comparable and to explore the performance of spontaneous speech data on prepared speech models further, there is a need for more spontaneous speech data that are transcribed in Standard German or comparable prepared speech data sets with Swiss German transcriptions.

6.2.3 Development of Mixed Models

The transcription of both data sets was different: The prepared speech data set was transcribed in Standard German, while the spontaneous speech data set was transcribed in Swiss German, following the Dieth rules. Not only was the comparability thereby weakened, but also extending a prepared speech model with spontaneous speech data was not possible. If there are some prepared and spontaneous speech data available with transcriptions in the same language, a model could be first fine-tuned on prepared speech data and then fine-tuned again on spontaneous speech data. It would be interesting to find out if one could increase a model's performance if one is trained on both speech styles.

6.2.4 Investigation of XLS-R

Whisper performed better on zero-shot learning. Therefore, Whisper was used to explore different improvement strategies on spontaneous speech data. However, the performance of XLS-R was proven to be better in the case of the comparison of the two models in both data sets. Therefore, it would be interesting to test the impact of improvement strategies on XLS-R as well.

6.2.5 Combinations of Parameters

In the scope of the thesis, only a few parameters could be adapted because retraining is time consuming. Future work could link and combine different parameters. The results showed that Whisper performs better on its default parameters, as suggested by the developers. The experiments to compare the default Whisper parameters with the version in which the parameters were adapted to XLS-R proved the importance of setting the correct hyperparameters. Therefore, improvement strategies could be tested on the basis of the default parameters. In addition, the number of steps between two checkpoints was suggested to be 1000. To avoid memory problems, the default parameters were tested on a number of saving steps of 100 instead of 1000. Therefore, the improvement strategies should be tested in a model version including the following parameters:

- Learning rate = 0.00001
- Total training batch size = 16
- Saving steps = 1000

It was out of scope to perform grid-search or other strategies to find the optimal set of hyperparameters in the case of XLS-R and Whisper. As the illustrations containing the training and validation loss over the given number of steps proved, fine-tuning a model for more steps is not always helpful. In many cases, the models start to overfit or stop improving. In the case of this thesis, early stopping was not applied as the goal was to make all developed models as comparable as possible and to investigate the behaviour if all models are trained for the defined 4000 steps. There is also evidence that adding dropout could help increase performance, as indicated by the Whisper model trained with disfluency labels and a dropout of 0.1. In addition to dropout, there are other strategies to avoid overfitting. Another example is adding noise to the audio, as clean audio restricts the ability of the model to generalise on unseen data. Some Gaussian noise was added to a spontaneous speech version of Whisper, but it did not increase the quality of the model. Therefore, this path was discarded. Future work could combine different of these proposed strategies.

6.2.6 Exploration of Dialects

There exist multiple dialects in Swiss German that differ in vocabulary, pronunciation, and syntax (Plüss et al., 2021) depending on the region. The performance of the ASR models in different dialects was shown by Schraner et al. (2022). Despite

other models, they also used a version of XLS-R and reported that the Innerschweiz dialect is the easiest, while some systems perform especially worse on the Wallis dialect. Based on these insights, it would be a step for future work to analyse the performance of spontaneous and prepared speech data on different dialects. It adds up to the data problem, as the Schawinski data consists only of speakers from the dialect region of Zurich. Therefore, one limitation is given by the difference in dialects. It is difficult to develop models that perform well in all different dialects. It was beyond the scope of this thesis to fine-tune models on different dialects or to prove performance differences if multiple dialects were used to test the model. Future work can tie in and further elaborate the approaches to achieve good results for all dialects.

6.2.7 Investigation of the Main Problems

In the scope of the thesis, it was not possible to analyse the predictions of the models on the test set in detail. Therefore, it is left for future work to analyse the behaviour of models on different part-of-speech tags, named entities, usage of foreign words, or even code-switching. Discovering error patterns in the two speech styles and comparing their occurrences could give insights for research in spontaneous speech in Swiss German.

6.2.8 Adding a Language Model

Another approach to improve the model, especially on spontaneous speech, would be an external language model to correct spelling errors and word order. The results in 4 show that spelling errors lead to a higher word error rate. Furthermore, the developers of the Whisper model explain spelling errors as an issue and a limitation of their models in their GitHub repository[1]. As described in Chapter 3, Whisper was pre-trained using some data scraped from the Internet. They used some filters to exclude machine-generated data. As they mention on their GitHub repository, these filters are not very effective for languages other than English. But they mention that a language model can be added to the Whisper decoder to improve the final output. It is not supported to include a language model, but the 'TokenDecoder' class can be extended to select tokens according to a language model.

[1] `https://github.com/openai/whisper/discussions/266`

Glossary

The glossary describes some terminology used in the scope of this thesis.

automatic speech recognition The process of **converting human speech** into a **machine-readable format** that makes it possible for humans to communicate with machines.

batch size Value to indicate the **number of samples** propagated through a model before the **parameters of the models** are updated. The **gradient accumulation step** can multiply the number of batches propagated before an update.

BLEU score A metric to **measure** the quality of a predicted text compared to the gold standard.

character error rate A metric to **evaluate the performance** of automatic speech recognition systems at the **character level**. The reference transcription and prediction of the systems are aligned, and the difference between the two sentences is calculated by the number of hits, substitutions, deletions, and insertions.

dropout A **regularisation** technique. It describes the **probability of dropping nodes** of the network to reduce overfitting. The dropout value can be set between 0 and 1.

evaluation Task that is performed during a training or fine-tuning process of a model after a given number of steps. Evaluation is carried out on the **validation set**.

epoch An **iteration** over the whole training data set.

gold standard A term used to describe an **annotation** in a data set that is collected and corrected by human annotators. The gold standard is validated to ensure the best possible annotation. This reference transcription is used to evaluate the **performance and quality** of speech recognition systems.

ground-truth A term used to describe an **annotation** in a data set that is collected by one or more human annotators. This reference transcription is used to evaluate the **performance and quality** of speech recognition systems.

gradient accumulation step Technique to **process more data** than the memory is capable of. It can cause slower processing of the data. The **gradients are accumulated** by the number of steps given in this parameter. Afterwards, the backward pass is performed. If the gradient accumulation step is set to 2, the per-device-batch-size is doubled if there are two GPUs available.

learning rate **Step size** in which a **minimum of the loss function** is approached.

optimiser Functions to **adjust the model parameters** during training to reduce the loss. AdamW[2] is an example of an optimiser. The optimiser consists of hyperparameters like beta and epsilon that can be set.

prepared speech A **speech style** characterised by being **carefully read** and being a formal speech. It serves as the opposite of spontaneous speech. The term describing this speech style can differ depending on the research paper.

speech-to-text The process of converting spoken language to text format. The speech-to-text system receives **spoken language** as an **audio signal** and outputs written text.

spontaneous speech A **speech style** characterised by **conversational and vernacular speech**. On the **contrary to the prepared speech**, spontaneous speech is not read and consists of fillers, hesitations, and incomplete sentences. The term describing this speech style can differ depending on the research paper.

step Evaluation strategy where the **performance of a model** is evaluated after a certain number of steps. There exists a **maximum number of steps** after which the training process is performed. The term evaluation steps refers to the number of steps between evaluations, while the term logging steps refers to the number of steps between two checkpoints.

tokeniser Used to **preprocess the input data** before training.

training loss Calculated **error** made by a machine learning model during training. Training loss indicates how well a model performs on the **training data**.

[2] `https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html`

validation loss Calculated **error** that a machine learning model makes during training. It indicates how well the model performs on **unseen data**. Is the validation loss higher than the training loss; the model is prone to overfitting as it cannot generalise on unseen data in the validation set.

word error rate A metric to **evaluate the performance** of automatic speech recognition systems at the **word level**. The reference transcription and prediction of the systems are aligned, and the difference between the two sentences is calculated by the number of hits, substitutions, deletions, and insertions.

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bundesamt für Statistik (2022). Hauptsprachen in der Schweiz. Ständige Wohnbevölkerung. `https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.assetdetail.21344032.html`. Accessed: 2023-02-05.

Dieth, E. and Schriftkommission, N. H. G. (1938). *Schwyzertütschi Dialäktschrift: Leitfaden [einer einheitlichen Schreibweise für alle Dialekte]*. Orell Füssli, Zürich, Schweiz.

Furui, S., Nakamura, M., Ichiba, T., and Iwano, K. (2005). Why Is the Recognition of Spontaneous Speech so Hard? In Matoušek, V., Mautner, P., and Pavelka, T., editors, *Text, Speech and Dialogue*, pages 9–22, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gabler, P., Geiger, B. C., Schuppler, B., and Kern, R. (2023). Reconsidering Read

and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information*, 14(2):137.

Graves, A. (2012). Sequence Transduction with Recurrent Neural Networks. *CoRR*, abs/1211.3711.

Gsteiger, M. and von Cranach, P. (2012). Dialekte. `https://hls-dhs-dss.ch/de/articles/024595/2012-04-19/`. Accessed: 2023-02-05.

Horii, K., Fukuda, M., Ohta, K., Nishimura, R., Ogawa, A., and Kitaoka, N. (2022). End-to-End Spontaneous Speech Recognition Using Disfluency Labeling. In *Proc. Interspeech 2022*, pages 4108–4112.

Huggingface (2023a). Huggingface Library: CER. `https://huggingface.co/spaces/evaluate-metric/cer`. Accessed: 2023-06-10.

Huggingface (2023b). Huggingface Library: Datasets. `https://huggingface.co/docs/datasets/index`. Accessed: 2023-06-01.

Huggingface (2023c). Huggingface Library: Datasets – Arrow. `https://huggingface.co/docs/datasets/about_arrow`. Accessed: 2023-06-01.

Juang, B. and Rabiner, L. (2005). Automatic Speech Recognition - A Brief History of the Technology Development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara.*

Li, J. (2021). Recent Advances in End-to-End Automatic Speech Recognition. *arXiv*. 10.48550/ARXIV.2111.01690.

Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *CoRR*, abs/2006.07264.

Morris, A. (2002). An Information Theoretic Measure of Sequence Recognition Performance. IDIAP. `http://infoscience.epfl.ch/record/82766`.

Morris, A. C., Maier, V., and Green, P. (2004). From WER and RIL to MER and WIL: Improved Evaluation Measures for Connected Speech Recognition. In *Proc. Interspeech 2004*, pages 2765–2768.

OpenAI (2023). Huggingface OpenAI: Whisper. `https://huggingface.co/openai/whisper-small`. Accessed: 2023-07-08.

Plüss, M., Deriu, J., Schraner, Y., Paonessa, C., Hartmann, J., Schmidt, L., Scheller, C., Hürlimann, M., Samardžić, T., Vogel, M., and Cieliebak, M. (2023). STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Plüss, M., Hürlimann, M., Cuny, M., Stöckli, A., Kapotis, N., Hartmann, J., Ulasik, M. A., Scheller, C., Schraner, Y., Jain, A., Deriu, J., Cieliebak, M., and Vogel, M. (2022). SDS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Plüss, M., Neukom, L., Vogel, M., and Christian, S. (2021). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Swiss Text Analytics Conference 2021*, Winterthur, Schweiz. `https://ceur-ws.org/Vol-2957/paper3.pdf`.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Scherrer, Y., Samardžić, T., and Glaser, E. (2019a). ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Scherrer, Y., Samardžić, T., and Glaser, E. (2019b). Digitising Swiss German: How to Process and Study a Polycentric Spoken Language. *Language Resources and Evaluation*, 53(4):735–769.

Schraner, Y., Scheller, C., Plüss, M., and Vogel, M. (2022). Swiss German Speech to Text System Evaluation. *arXiv*. 10.48550/ARXIV.2207.00412.

Tsvetkov, Y. (2017). *Opportunities and Challenges in Working with Low-Resource Languages*. Language Technologies Institute. Carnegie Mellon University. `https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf`.

Tucker, B. V. and Mukai, Y. (2023). *Spontaneous Speech*. Elements in Phonetics. Cambridge University Press, Cambridge.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wang, P., Sun, R., Zhao, H., and Yu, K. (2013). A New Word Language Model Evaluation Metric for Character Based Languages. In Sun, M., Zhang, M., Lin, D., and Wang, H., editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 315–324, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yadav, H. and Sitaram, S. (2022). A Survey of Multilingual Models for Automatic Speech Recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

A Access to the Code and Models

A.1 Accessing the Code

The code can be found in the following GitHub repository: `https://github.com/KarinTho/master-thesis-STT` This includes the code for preprocessing, the training of the models, and the evaluation. Furthermore, the output of the evaluation is also saved on GitHub.

A.2 Accessing the Models

Some models were saved in the Huggingface Hub repository: `https://huggingface.co/karinthommen`.

The following table provides an overview of the models, in which coding file they can be found, and the model's name as it is saved in the Huggingface repository. The model can be loaded by adding the repository name to the following code:

$$xlsr = Wav2Vec2ForCTC.from\_pretrained(repo\_name)$$

or

$$whisper = WhisperForConditionalGeneration.from\_pretrained(repo\_name)$$

with $repo\_name = Model\ name\ on\ Huggingface$

Model	Speech Type	Code	Model Name on Huggingface
Comparable XLS-R	prepared speech	STT_1_XLSR_V3.ipynb Version 3.3	karinthommen/xlsr-prep-small-2
Large XLS-R	prepared speech	STT_1_XLSR_V3.ipynb Version 3.1	karinthommen/xlsr-V3-2
Comparable Whisper	prepared speech	STT_2_Whisper_V4.ipynb Version 4.4	karinthommen/whisper-V4-small-3
Large Whisper	prepared speech	STT_2_Whisper_V4.ipynb Version 4.1	karinthommen/whisper-V4-2
Default Whisper Large	prepared speech	STT_2_Whisper_V2.ipynb Version 2.1	karinthommen/whisper-V2
Default Whisper Small	prepared speech	STT_2_Whisper_V2.ipynb Version 2.3	karinthommen/whisper-V2-default-small
German Whisper	prepared speech	STT_2_Whisper_V4.ipynb Version 4.2	karinthommen/whisper-V4-small
Comparable XLS-R	spontaneous speech	STT_4_Spont_XLSR_V1.ipynb Version 1.2	karinthommen/spont-xlsr-V1-2
Comparable Whisper	spontaneous speech	STT_3_Spont_Whisper_V2_1.ipynb Version 2.1	karinthommen/spontaneous-whisper-v2-4
Default Whisper	spontaneous speech	STT_3_Spont_Whisper_V2_2.ipynb Version 2.2	karinthommen/spont-whisper-default
German Whisper	spontaneous speech	STT_3_Spont_Whisper_V4_2.ipynb Version 4.2	karinthommen/spontaneous-whisper-v4-2
Punctuation Whisper	spontaneous speech	STT_3_Spont_Whisper_V5.ipynb Version 5.4	karinthommen/spontaneous-whisper-v5-4
Disfluency Whisper	spontaneous speech	STT_3_Spont_Whisper_V6.ipynb Version 6.1	karinthommen/spontaneous-whisper-v6
Disfluency and Dropout	spontaneous speech	STT_3_Spont_Whisper_V6.ipynb Version 6.3	karinthommen/spontaneous-whisper-v6-3

Table A.1: Overview of the code and models.

Access to the models that were not used in the scope of this thesis but visible in the coding files can be requested by writing an email to karin.thommen@uzh.ch.

A.3 Overview of the Output of the Models on the Test Data

The files mentioned in the table can be found on GitHub: `https://github.com/KarinTho/master-thesis-STT/tree/main/Evaluation`. The table provides an overview of which model correspond to which evaluation file. The corresponding testing script is also available on GitHub: `https://github.com/KarinTho/master-thesis-STT/blob/main/testing.ipynb`

Model	Trained Speech Type	Tested Speech Type	Filename Model Output
Comparable XLS-R	prepared speech	prepared speech	test_xlsr_prep_small_on_prep.csv
		spontaneous speech	test_xlsr_prep_small_on_spont.csv
Large XLS-R	prepared speech	prepared speech	test_xlsr_v3_2_prep.csv
		spontaneous speech	test_xlsr_v3_2_spont.csv
Comparable Whisper	prepared speech	prepared speech	test_whisper_small_prep_on_prep.csv
		spontaneous speech	test_whisper_small_prep_on_spont-2.csv
Large Whisper	prepared speech	prepared speech	test_whisper_v4_prep.csv
		spontaneous speech	test_whisper_v4_spont.csv
Default Whisper Large	prepared speech	prepared speech	test_whisper_v2_prep.csv
		spontaneous speech	test_whisper_v2_prep_on_spont.csv
Default Whisper Small	prepared speech	prepared speech	test_whisper_v2_small_prep_on_prep.csv
		spontaneous speech	test_whisper_v2_small_prep_on_spont.csv
German Whisper	prepared speech	prepared speech	test_whisper_small_prep_on_prep_german.csv
Comparable XLS-R	spontaneous speech	spontaneous speech	test_xlsr_spont_on_spont.csv
		prepared speech	test_xlsr_spont_on_prep.csv
Comparable Whisper	spontaneous speech	spontaneous speech	test_spont_whisper_on_spont.csv
		prepared speech	test_spont_whisper_on_prep.csv
Default Whisper	spontaneous speech	spontaneous speech	test_spont_default_whisper_on_spont.csv
		prepared speech	test_spont_default_whisper_on_prep.csv
German Whisper	spontaneous speech	spontaneous speech	test_spont_whisper_v4_2.csv
Punctuation Whisper	spontaneous speech	spontaneous speech	test_spont_whisper_v5_4.csv
Disfluency Whisper	spontaneous speech	spontaneous speech	test_spont_whisper_v6.csv
Disfluency and Dropout	spontaneous speech	spontaneous speech	test_spont_whisper_v6_3.csv
Zero Shot XLS-R	-	prepared speech	test_xlsr_zero_prep.csv
		spontaneous speech	test_xlsr_zero_spont.csv
Zero Shot Whisper	-	prepared speech	test_whisper_zero_prep.csv
		spontaneous speech	test_spont_whisper_zero.csv

Table A.2: Overview of the outputted prediction by the models on the test data.

B Report of the Models

This part of the appendix contains some information about the models, sample predictions and test results.

B.1 Comparable Models

B.1.1 Comparable Wav2Vec2 XLS-R

Comparable XLS-R							
Prepared Speech (XLS-R-small)				Spontaneous Speech			
step	**train loss**	**validation loss**	**WER**	**step**	**train loss**	**val loss**	**WER**
400	3.42	1.72	97.27	**400**	3.23	1.51	97.95
800	0.84	1.93	87.17	**800**	0.95	0.93	71.93
1200	0.26	2.29	85.02	**1200**	0.41	1.06	67.24
1600	0.15	2.40	82.83	**1600**	0.26	1.14	64.61
2000	0.11	2.47	80.33	**2000**	0.19	1.15	64.74
2400	0.08	2.62	80.11	**2400**	0.15	1.19	63.76
2800	0.06	2.62	79.44	**2800**	0.11	1.25	62.78
3200	0.05	2.70	79.75	**3200**	0.09	1.25	60.64
3600	0.04	2.66	77.96	**3600**	0.07	1.28	60.44
4000	0.03	2.71	78.41	**4000**	0.06	1.29	59.66

Table B.1: Training loss, validation loss and word error rate for the saved training steps for the comparable XLS-R models. The word error rate is reported in percentage.

B.1.2 Comparable Whisper

Comparable Whisper							
Prepared Speech				Spontaneous Speech			
step	**train loss**	**validation loss**	**WER**	**step**	**train loss**	**val loss**	**WER**
400	0.97	3.07	88.80	**400**	0.71	2.10	76.65
800	0.20	3.66	91.98	**800**	0.18	2.34	66.92
1200	0.11	3.99	88.13	**1200**	0.09	2.52	70.40
1600	0.07	4.23	91.08	**1600**	0.06	2.67	70.37
2000	0.04	4.27	87.59	**2000**	0.04	2.74	70.50
2400	0.02	4.23	86.38	**2400**	0.02	2.57	66.75
2800	0.01	4.13	84.95	**2800**	0.00	2.60	66.30
3200	0.00	4.13	85.44	**3200**	0.00	2.65	65.22
3600	0.00	4.14	84.95	**3600**	0.00	2.69	65.87
4000	0.00	4.14	84.99	**4000**	0.00	2.70	66.13

Table B.2: Training loss, validation loss and word error rate for the saved training steps for the comparable Whisper models. The word error rate is reported in percentage.

B.2 Impact of the Data Set Size

Large XLS-R				Large Whisper			
Prepared Speech				Prepared Speech			
step	**train loss**	**validation loss**	**WER**	**step**	**train loss**	**validation loss**	**WER**
400	3.52	1.70	96.96	**400**	1.76	1.58	90.86
800	1.29	1.15	69.99	**800**	1.34	1.31	76.83
1200	0.99	1.02	60.63	**1200**	1.09	1.10	87.38
1600	0.88	0.91	54.23	**1600**	0.93	0.98	79.38
2000	0.84	0.86	50.51	**2000**	0.82	0.88	73.18
2400	0.78	0.82	47.16	**2400**	0.71	0.77	79.21
2800	0.75	0.78	45.09	**2800**	0.55	0.72	72.17
3200	0.75	0.76	43.69	**3200**	0.36	0.66	69.82
3600	0.68	0.73	42.75	**3600**	0.33	0.61	63.11
4000	0.61	0.71	41.66	**4000**	0.29	0.58	61.55

Table B.3: Training loss, validation loss and word error rate for the saved training steps for the large XLS-R and Whisper models. The word error rate is reported in percentage.

B.3 Impact of the Default Parameters

Default Whisper Large				Default Whisper Small				Default Whisper			
Prepared speech				Prepared Speech				Spontaneous Speech			
step	train loss	val loss	WER	step	train loss	val loss	WER	step	train loss	val loss	WER
100	0.727	0.798	45.892	**100**	3.555	1.733	72.106	**100**	2.617	1.282	61.191
200	0.594	0.724	45.448	**200**	1.008	0.864	727.381	**200**	0.892	0.994	51.366
300	0.596	0.685	47.272	**300**	0.352	0.843	1038.445	**300**	0.524	0.938	48.504
400	0.591	0.660	47.704	**400**	0.197	0.869	869.329	**400**	0.312	0.948	47.495
500	0.583	0.650	48.230	**500**	0.117	0.863	1230.073	**500**	0.206	0.945	46.942

Table B.4: Training loss, validation loss and word error rate for the saved training steps for the default Whisper models. The word error rate is reported in percentage.

B.4 Improvement Strategies

B.4.1 Adding a Language Token

Comparable Whisper with German token				Whisper with German token			
Prepared Speech				Spontaneous Speech			
step	train loss	validation loss	WER	Step	train loss	validation loss	WER
400	0.83	2.95	84.67	**400**	0.96	2.15	77.63
800	0.17	3.38	89.91	**800**	0.18	2.41	90.72
1200	0.09	3.85	98.58	**1200**	0.09	2.65	85.61
1600	0.06	3.86	96.58	**1600**	0.05	2.69	95.47
2000	0.03	3.97	92.89	**2000**	0.03	2.81	94.27
2400	0.02	3.97	95.07	**2400**	0.01	2.67	97.43
2800	0.01	3.94	93.56	**2800**	0.00	2.72	95.28
3200	0.00	3.92	93.47	**3200**	0.00	2.77	94.53
3600	0.00	3.93	93.65	**3600**	0.00	2.79	94.66
4000	0.00	3.93	93.43	**4000**	0.00	2.79	94.69

Table B.5: Training loss, validation loss and word error rate for the saved training steps for the Whisper models containing the German language token. The word error rate is reported in percentage.

B.4.2 Keep Punctuation

Punctuation Whisper			
Spontaneous Speech			
Step	**train loss**	**validation loss**	**WER**
400	0.69	1.98	68.84
800	0.16	2.31	76.98
1200	0.09	2.44	75.55
1600	0.06	2.67	73.98
2000	0.04	2.58	72.09
2400	0.02	2.57	71.08
2800	0.00	2.58	69.13
3200	0.00	2.63	67.99
3600	0.00	2.65	68.06
4000	0.00	2.66	67.93

Table B.6: Training loss, validation loss and word error rate for the saved training steps for the Whisper where the punctuation was kept in the training data. The word error rate is reported in percentage.

B.4.3 Disfluency Labelling

Whisper: Disfluency Labelling				Whisper: Disfluency Labelling and Dropout = 0.1			
Spontaneous Speech				Spontaneous Speech			
step	**train loss**	**val loss**	**WER**	**step**	**train loss**	**validation loss**	**WER**
400	0.72	1.76	100.00	**400**	0.72	1.76	60.82
800	0.16	2.14	85.24	**800**	0.16	2.14	69.06
1200	0.08	2.22	75.52	**1200**	0.08	2.22	65.33
1600	0.05	2.31	83.24	**1600**	0.05	2.31	63.73
200	0.03	2.32	80.24	**2000**	0.03	2.32	63.82
2400	0.01	2.33	87.97	**2400**	0.01	2.33	62.91
2800	0.00	2.32	83.82	**2800**	0.00	2.32	62.30
3200	0.00	2.38	80.58	**3200**	0.00	2.38	61.97
3600	0.00	2.40	80.12	**3600**	0.00	2.40	61.55
4000	0.00	2.41	79.97	**4000**	0.00	2.41	61.24

Table B.7: Training loss, validation loss and word error rate for the saved training steps for the Whisper models implemented with disfluency labelling. The word error rate is reported in percentage.

B.5 Applicability

This table shows the results of the comparable models if they are tested on the opposite data set and not only on the one on which they are trained.

Comparable Whisper							
Trained on Prepared Speech				**Trained on Spontaneous Speech**			
Tested with Prepared Speech		**Tested with Spontaneous Speech**		**Tested with Spontaneous Speech**		**Tested with Prep. Speech**	
WER	91.89	**WER**	101.72	**WER**	67.02	**WER**	115.81
CER	64.64	**CER**	77.29	**CER**	33	**CER**	64.5
BLEU	0.045	**BLEU**	0	**BLEU**	0.156	**BLEU**	0.001

Table B.8: Comparable Whisper trained on prepared speech or spontaneous speech data and tested on the corresponding and opposite test set.