Zurich University
of Applied Sciences

School of
Engineering

Project Thesis

ZHAW School of Engineering

CAI, Centre for Artificial Intelligence

Automatic Detection of Swiss German Dialects using Wav2Vec

Authors:
Samuel Stucki
Patrik Randjelovic

Supervisor:
Prof. Dr. Mark Cieliebak

Secondary Supervisor:
Jan Deriu

December 24, 2021

Abstract

Swiss German is a collection of Alemannic dialects spoken mostly by the German-speaking cantons of Switzerland. In contrast to other dialects, Swiss German is to this day an important part of the Swiss identity and thus spoken in practically all social settings. The low population size of Switzerland and its diverse dialects make it difficult to collect sufficient data for the task of speech recognition and translation. This thesis aims at applying a new Swiss German corpus on the low-resource specialized multilingual Wav2Vec-XLSR Transformer architecture for Dialect Identification (DID) in the domain of Automatic Speech Recognition (ASR). Previous research on low-resource languages showed promising results when applied to the Wav2Vec architecture and the solution proposed in this work builds upon these findings. Multiple granularities were tested to gain insight into the relations of the dialects to each other. The best results were achieved when the dialects were grouped into four distinct regions based on their linguistic similarity as well as their geographic proximity. The best region, the eastern High-Alemannic group, achieved an F1-score of 0.65 while the model as a whole reached a weighted F1-score of 0.5. Based on these encouraging first findings future research should be able to improve the performance and develop a well-functioning ASR system based on the Wav2Vec architecture for Swiss German.

Zusammenfassung

Schweizerdeutsch ist eine Sammlung von alemannischen Dialekten, welche vor allem in den deutschsprachigen Kantonen der Schweiz gesprochen werden. Im Gegensatz zu anderen Dialekten ist Schweizerdeutsch bis heute ein wichtiger Teil der Schweizer Identität und wird daher in praktisch allen gesellschaftlichen Bereichen gesprochen. Die geringe Einwohnerzahl der Schweiz und ihre vielfältigen Dialekte machen es schwierig, genügend Daten für Spracherkennung und Übersetzung zu sammeln. Diese Arbeit zielt darauf ab, einen neuen schweizerdeutschen Korpus auf die low-resource spezialisierte mehrsprachige Wav2Vec-XLSR Transformer-Architektur zur Dialektidentifikation (DID) im Bereich der automatischen Spracherkennung (ASR) anzuwenden. Frühere Untersuchungen im Bereich der ressourcenarmen Sprachen zeigten vielversprechende Ergebnisse, wenn sie auf die Wav2Vec-Architektur angewandt wurden. Die in dieser Arbeit vorgeschlagene Lösung baut auf diesen Ergebnissen auf. Mehrere Granularitäten wurden getestet, um einen besseren Einblick in die individuellen Beziehungen der Dialekte untereinander zu erhalten. Die besten Ergebnisse wurden erzielt, als die Dialekte basierend auf ihrer sprachlichen Ähnlichkeit und ihrer geografischen Nähe in vier verschiedene Regionen eingeteilt wurden. Die beste Region, die östliche hochalemannische Gruppe, erreichte einen F1-score von 0,65, während das Modell als Ganzes einen gewichteten F1-score von 50,23% erreichte. Auf der Grundlage dieser ermutigenden ersten Ergebnisse sollten zukünftige Arbeiten in der Lage sein, die Leistung zu verbessern und ein gut funktionierendes ASR-System für Schweizerdeutsch zu entwickeln.

Preface

This work has given us the opportunity to delve into certain topics that would otherwise only be touched on during the lectures. It also allowed us to discover and work with the latest models and concepts in ASR such as Wav2Vec. Systems that exploit AI to perform complex actions are a topic that has always been fascinating, and when the opportunity came up to get involved, we were happy to do so. This thesis is aimed at all those who have an interest in ASR systems and are looking to understand their potential as well as see how and with what tools one can classify the various Swiss-German dialects.

We would like to thank Prof. Dr. Mark Cieliebak and Jan Deriu for all their invaluable support during this semester. We also thank Pascal Fivian and Dominique Reiser, who composed the Bachelor thesis, of which this thesis is a continuation, for their advice.

Contents

Chapter 1

Introduction

Switzerland has four official languages: German, French, Italian, and Romansh. German (including dialects) is the most prevalent of the four languages with more than 60% of the population speaking it [1] and is adopted by 21 of the total 26 cantons. Standard German is used in official settings and used for general understanding while Swiss German, called "Schwyzerdütsch" in the native tongue, is used for everyday life. Swiss German consists of a set of dialects derived mainly from Alemannic German. The relatively small size of the cantons and the lack of an orthographic standard has led to a wide variety of dialects that are difficult to classify. [2]

Neural network models traditionally yield good results when they are fed large amounts of labelled data. The application for Swiss German in this context is limited because the currently available amount of labelled data is very low. It is also obvious that it will be difficult to significantly increase the data size in the future compared to other, more widespread languages like German which has around 135 million people speaking it [3], against Swiss German with approximately 5 million speakers.

The latest research aimed at transcribing the various dialects of Swiss German has focused on a single, specifically tuned ASR model obtaining positive results. However, it remains considerably behind models trained to recognise e.g. English. This is mainly due to the lack of available data. [4]

We believe that better results can be achieved by training multiple ASR models for each dialect or a group of similar dialects using the state-of-the-art self-supervised learning model Wav2Vec 2.0 XLSR. Therefore, the first step, which this work aims at, is to recognise them autonomously based on the audio tracks to subsequently use the appropriate ASR model. Previous work using this model suggest promising results in identifying different accents and languages in the context of Cross-Lingual transfer learning.[5][6]
For this project, we analyze the data used and perform experiments to fine-tune and evaluate the potential of this model.

1.1 Literature review

In Natural Language Processing (NLP) Automatic Speech Recognition (ASR) is an important task in general and thus a strongly researched topic. The underlying subjects of Accent Identification (AID) and Dialect Identification (DID) have also become big topics in the research community. Over the last decade, a variety of different approaches have been made to solve this complex challenge.

A paper by Zaidan and Callison-Burch from 2014 tried classifying Arabic dialects, specifically those of the Levante, Gulf, and Egypt, against Modern Standard Arabic (MSA) using smoothed n-gram models in multiple experiments with dialect-annotated data from different newspapers. First, a two-way classifier was used, where the dialects for each of the three regions were put against MSA. An accuracy of 85.7% was achieved with a word-based unigram classifier. After using a multi-way classification, including MSA, the accuracy dropped to 81.0%. When they removed MSA from the classes the accuracy increased to 88.4% however, which was higher than the best two-way classifier from the first experiments. This indicated that the dialects are more distinguishable from each other than from MSA itself. [7]

In 2017 and 2018 Zampieri et al. organized a Varieties and Dialects (VarDial) workshop at the European Chapter of the Association for Computational Linguistics (EACL) Conference. Included in this workshop were two tasks concerning the identification of either five Arabic or four Swiss German dialects. [8] The best result on the Arabic DID task was in 2017 and used machine learning with multiple kernel learning based on i-vectors and achieved an accuracy of 76.27% and a weighted F1-Score of 76.32%. [9] The task for the Swiss German DID in 2018 resulted in a macro F1-Score of 68.6% where a Word-Based Backoff method was used for the identification. [10]

Weninger et al. used Deep Learning in 2019 for classifying 15 regional mandarin accents which resulted in an accuracy of up to 34.5%. After they clustered regionally related accents into three distinct groups, they were able to improve the accuracy and achieved an unweighted average recall of 66.4% with a bidirectional Long Short-Term Memory (bLSTM) classifier. [11]

There have been several works on Spanish dialects as well. One of them used two on dialects retrained Gaussian Mixture Models (GMM) called Mixture-Selection GMM (MS-GMM) and Frame Selection GMM (FS-GMM). Applying these models to 3.3h of speech from three dialect areas, in this case, Cuba, Peru, and Puerto Rico showed that they outperformed the baseline GMM significantly. While the GMM achieved 74.3% accuracy, the MS-GMM and FS-GMM attained an accuracy of 82.1% and 81.0% respectively. [12]

The Nuanced Arabic Dialect Identification (NADI) shared task in 2021 had two subtasks concerning the identification of Arabic dialects on either country- or province-level based on a dataset of publicly available Twitter messages. [13] They targeted

21 Arabic countries with a combined amount of 100 provinces. The best team won all given tasks by using a specifically on Arabic tailored version of the pre-trained transformer language model BERT [14] called MARBERT [15]. In the country-level subtask, an accuracy of 51.66% and a macro F1-Score of 32.26% were achieved. On the second subtask, where contestants had to identify the province based on the dialect in the Tweets, an accuracy of 9.46% and an 8.60% macro F1-Score was accomplished, thus showing that distinguishing a dialect on a small granularity like a province or state is possible but challenging. This also showcases the emergence of self-supervised learning neural networks such as BERT or Wav2Vec from Facebook's AI Team.

By classifying accents in Spanish and English in [5] it was shown that Wav2Vec-XLSR-53 is capable to distinguish accents from each other. By testing different lengths of audio samples they were also able to prove that longer speech samples of 5 to 10 seconds perform better than classifications than shorter samples. Though no dialect identifications were performed in their experiments, a proposal was given by the team to group similar dialects together as they should improve the performance of the Wav2Vec model.

1.2 Outline

This thesis is structured as follows. After the Introduction in Chapter 1, which also includes a Literature Review, there is the Foundation Chapter 2, containing the essential tools for understanding this work. This is followed by the Data Collection Chapter 3, where the data from the corpus provided by SwissNLP are analysed and subsequently used for the various experiments. In the Experimental Setup Chapter 4 an in-depth outline of the performed experiments is given. The obtained results with the individual conclusions can be found in the Results Chapter 5. Finally, in Discussion and Outlook Chapter 6, an overarching conclusion of the work is made taking into account the knowledge gained throughout this thesis. A possible continuation of this topic is proposed as well.

Chapter 2

Foundations

2.1 Speech Processing

The idea of designing a machine that can mimic the human ability of speaking has been a topic of interest since the late 19th century. [16] According to Shannon's information theory [17] speech can be encoded, transmitted, manipulated, recorded, and lastly decoded by a receiver via a continuously varying waveform. This concept was first implemented in 1952 by the famous Bell telephone. [16] Since then various inventions like the Markov model [18] and specializations in the field itself such as Speech Recognition and Text-to-Speech have helped the field progress further. Now with the advent of deep learning in the last decade, a transition from traditional ASR models that are largely based on GMM's to neural networks like CNN's is happening.

2.2 Dialect Identification

Dialect identification (DID) refers to the attempt at recognizing dialects in an automated manner. Dialects are varieties of a particular language that differ in grammar, pronunciation, or vocabulary. [19] While dialects are usually regional there also exist class and occupation-specific dialects which give insight into a person's social background or occupation the individual performs. Swiss German is a collection of regional dialects and as such will be the only considered variation. DID is one of the most complex aspects of speech recognition because these dialects are specific to a regional community and not standardized in any way. Furthermore, they are more susceptible to change over time than normal languages. This is even further impeded by the general lack of text and audio data. [20] This results in a need for models that can give accurate predictions on low resource data.

2.3 Wav2Vec

Wav2Vec was first released in 2019 by Facebook's AI team to remove the dependency on vast amounts of training data that would traditionally be needed for a model to yield practical results in speech recognition as data collection for uncommon languages and dialects such as Swiss German is not feasible most of the time. The model is a convolutional neural network and made use of unsupervised pre-training on large amounts of unlabeled audio data. Training is done in two phases consisting of a pre-training and a second fine-tuning phase. First test results on the WSJ and nov92 test sets all showed significant increases in performance on magnitude less of labeled data. [21]

After further research subsequent models were published, beginning with vq-Wav2Vec [22] and in 2020 the second generation Wav2Vec 2.0 [23] was released. One significant change in these models was that instead of applying unsupervised learning they switched to self-supervised learning which does not focus on clustering and grouping. The model notably improved previous scores set by the first generation. Parallel to the base model, which was exclusively pre-trained on English data, a cross-lingual model named Wav2Vec2-XLSR-53 that had been pre-trained on 53 different languages was published in the hopes of improving research on low-resource languages. [24]

During the writing of this thesis an even newer model, Wav2Vec2-XLS-R, has been released which allegedly seriously improved upon the XLSR-53 model. [25] This thesis is still based on applying the XLSR-53 model to a Swiss German dataset with the newest XLS-R model being a factor of interest for further studies.

For a better understanding of Wav2Vec's inner workings, an introduction to each important component has to be given.

2.3.1 Contrastive Learning

As already explained, collecting labelled data is often difficult for a lot of real-world scenarios as correct labelling must be performed by experienced professionals that need to spend countless hours looking at images or listening to audio files. Here is where contrastive learning can help. In [26] it is explained that contrastive learning, in its essence, learns by comparison without knowing what the information it compares stands for. Unlike other methods where individual data samples are taken one at a time to learn a signal from them, this method learns by comparing among the samples themselves. The goal of this approach is to cluster "similar" and "dissimilar" samples together while keeping them at a distance from each other. One advantage of this approach is that the model architecture does not have to be modified between training and fine-tuning. Wav2Vec uses this technique to learn latent speech representation from raw audio data. [23]

2.3.2 Latent Speech Representation

In linguistics, phonemes are a concept describing the smallest unit of speech. They allow a speaker of a particular language to distinguish words or word elements from one another. Phonemes are uniquely classified for each spoken language. While English may have one way of pronouncing the element "p" Thai has two. [27]

The theoretical number of phonemes in any language is low, in speech waveforms, however, which contain both acoustic and linguistic information, these can increase dramatically because of individual speaking style, dialect, emotional state, environments, and much more. [28] Research tried capturing these influences with handcrafted features for a long time and only in recent years has unsupervised learning of these factors been tried. The application of this automatic approach in transformer models such as BERT or Wav2Vec has yielded significant improvements in speech recognition in general. [14] [23] Latent speech representation thus refers to the attempt of inferring latent variables from these speech waveforms in form of phonemes and representing them in a latent space.

2.3.3 Transformers

In 2017 Vaswani et al. [29] proposed a new model architecture, called Transformers, that aimed to remove the constraints hindering existing state-of-the-art sequence-to-sequence Recurrent Neural Networks (RNN). RNNs are by design sequential in nature which makes them unable to be used in a parallel environment. This has impacts on factors such as memory or training duration when longer sequence lengths, like in speech processing, are applied. The Transformer architecture solves this by entirely relying on a self-attention mechanism. This attention method allows the model for every single sequence in an input to look at the rest of the input and thus better understand the relation or role that a specific sequence has within that input. This self-attention is then applied multiple times in so-called encoders and decoders. These two components are nearly identical to each other, the key difference is that the decoder also takes input from the encoder as well as feeding its output back to itself after each step. Figure 2.1 shows the Transformer architecture on the left side and an abstracted version on right.

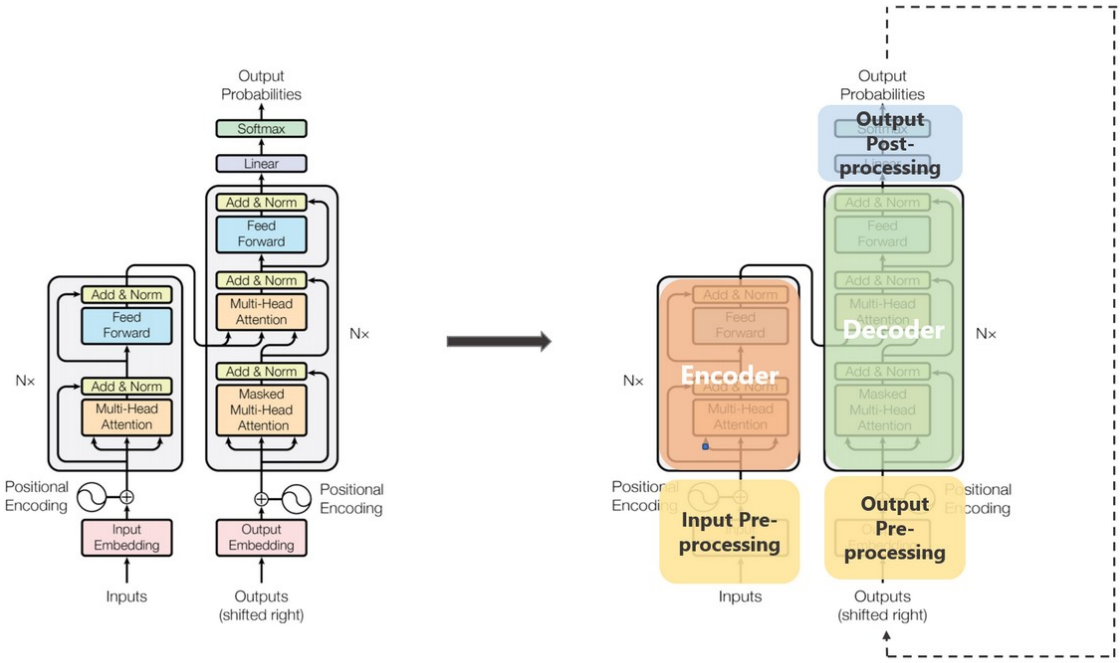

Figure 2.1: Architecture of Transformers, figure taken from [29] [30]

Each of these abstracted components in Figure 2.1 is explained in more detail below.

Self-Attention

Self-attention is a multi-step mechanism which we try to explain here by using examples from [29] and [31]. The first step concerns the input and output pre-processing which are not part of the self-attention mechanism but are prerequisites for it to function properly. First, the processing text is cut into pieces called Tokens and many variations exist on how to do this. For better visualization one can best imagine these Tokens by splitting a simple sentence like "A walk by the river bank." into its words and special characters. These Tokens are then encoded into an embedding of real, continuous values in form of a vector. A Token, regardless of its context, will always have the same embedding thus giving computers the ability to compare.

The mechanism then calculates the scalar product of two embeddings to determine how related the Tokens are to each other. Embeddings of words like bank and river are similar to each other because they both encode the same aspect of nature. The resulting value is consequently higher for related embeddings than unrelated. However, just calculating the scalar for each embedding would not provide any important information other than the relation about two Tokens in a sentence. For the device to understand grammatical features three distinct projection vectors, namely keys K, queries Q, and values V, must be made. Each projection is also mapping the input embedding on a lower dimension and represents a specific semantic aspect like a preposition, location, and place. This way the scalar products between the key and query can be focused and can represent relevant relationships. Important

to note is that these aspects are not fixed and the model is free to decide which combinations are most helpful with the given task at hand.

The resulting scalar products are then passed through an activation function, in the original case softmax, to make large relationships exponentially more significant and move smaller or negative ones towards zero. Additionally, normalization is performed as well so that each column sums up to one. Lastly, a contextualized embedding vector is created by a linear combination of all corresponding value vectors V with the proportions given by the softmax.

Figure 2.2 shows the concepts explained above.

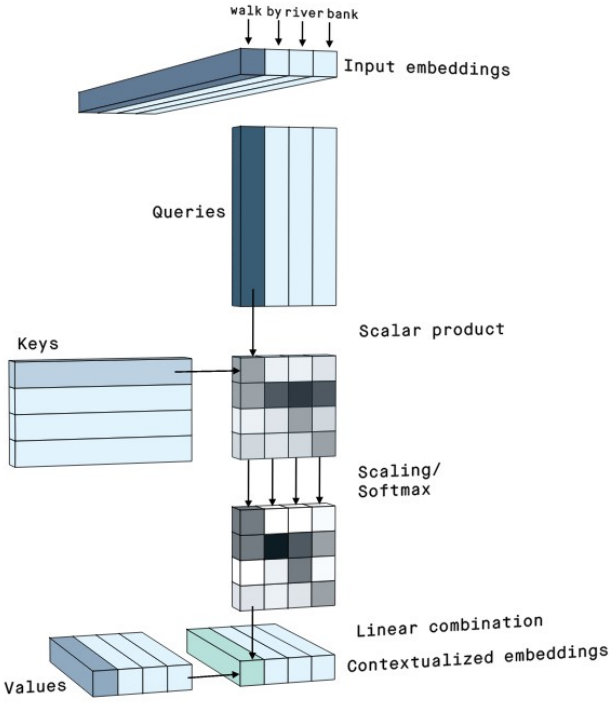

Figure 2.2: Mechanism of self-attention, figure taken from [31]

Multi-head attention

The process explained in self-attention can also be repeated multiple times for the same input embedding using different key, query, and value projections, forming what is called Multi-head attention as visualized in Figure 2.3. Each head can focus on different types of relationships between the tokens creating specific outputs. These outputs of the respective heads are then concatenated to one large contextualized embedding. [31]

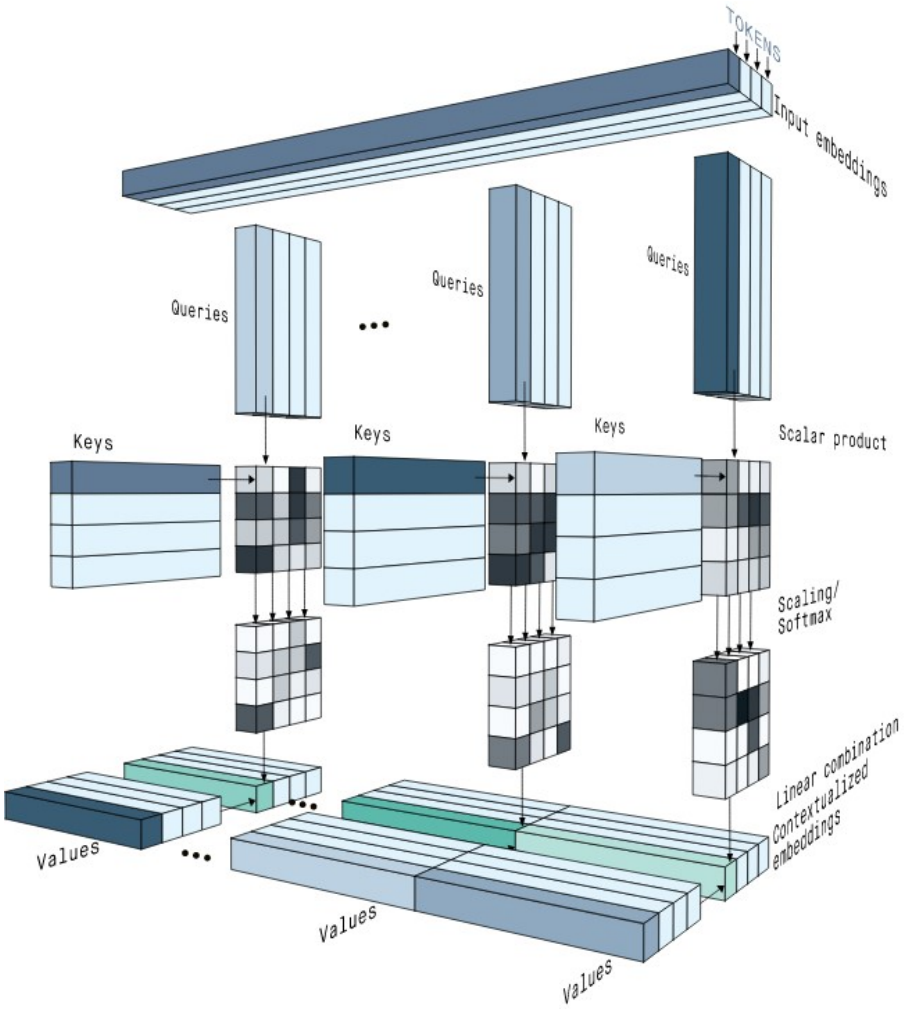

Figure 2.3: Architecture of Multi-head attention, figure taken from [31]

Encoder and Decoder

As explained in [29] both the encoder and decoder consist of a stack of identical layers. The Encoder layers each have two sublayers consisting of the multi-head attention and a position-wise fully connected feed-forward network which is applied to each position separately and identically. The decoder has one more sublayer compared to the encoder which is a second multi-head attention layer. This attention layer performs its operation over the output of the encoder. Additionally, a masked multi-head attention layer is applied to the output of the decoder to prevent positions from attending subsequent positions. Important to note is the positional encoding that must be applied to the input embedding as the model has no recurrence and convolution and would otherwise be unable to make use of the order of tokens in a sequence.

2.3.4 Quantization

The process of quantization converts values from a continuous space like latent speech representation into a discrete space. To achieve this Wav2Vec provides G codebooks with V entries where the best entry from each codebook is chosen for every latent speech representation z_t. All these entries are then concatenated into a vector e_t on which a linear transformation is applied to obtain q_t. [23] The process is visualized in Figure 2.4.

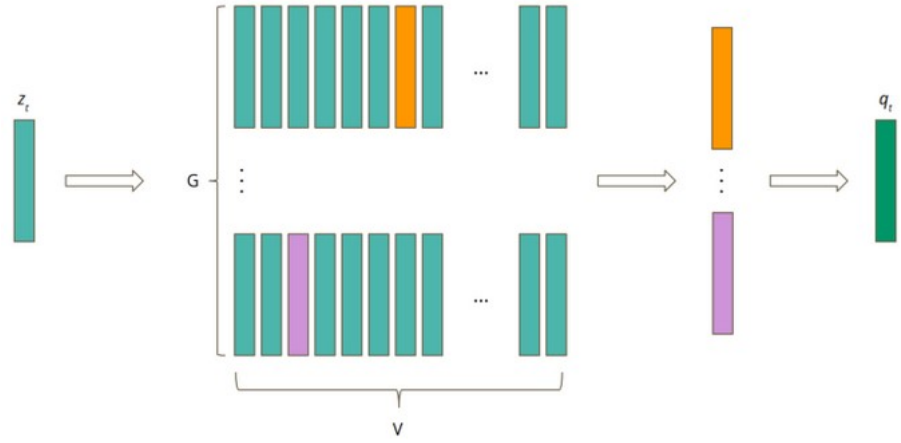

Figure 2.4: Quantization process, figure taken from [32]

Wav2Vec2 used the Gumbel softmax in this step as the addition of randomization allows the model to choose different code words more freely during the early stages of training and update their weights accordingly. The usage of so-called temperature is then applied to reduce the impact of randomization over time.[23]

2.3.5 Wav2Vec 2.0 Model Architecture

The Wav2Vec model is a combination of all the above-discussed topics as seen in Figure 2.5. The model is given a raw waveform X that is fed into a multi-layer convolutional feature encoder $f: X \mapsto Z$ which they call a feature extractor and yields latent speech representations $z_1, \ldots, z_T$ for T time-steps. These latent speech representations are then given to a Transformer $g: Z \mapsto C$ to build representations $c_1, \ldots, c_T$. At the same time, the output of the feature encoder is discretized into q_t that represent the targets of the objective. [23]

The datasets used in Wav2Vec 2.0 for pre-training were the Librispeech corpus containing 960 hours of audio and the LibriVox corpus with 53.2k hours of audio. The pre-training of the model was done similarly to the masked language model BERT [14] by randomly selecting parts of the latent speech representation and masking them before feeding them into the Transformer. [23]

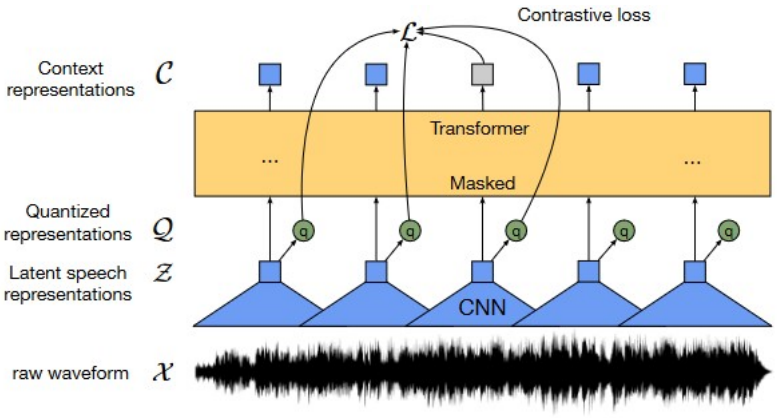

Figure 2.5: Wav2Vec 2.0 model architecture, figure taken from [23]

2.3.6 Wav2Vec2 XLSR Model Architecture

The Wav2Vec2-XLSR-53 cross-lingual model is still largely based on the Wav2Vec 2.0 proposed by Baevski et al. [23] but instead of just using English data it uses datasets that contain 53 different languages. The goal of this approach was for the model to learn representations of speech generalized across different languages which can be seen in Figure 2.6. [24]

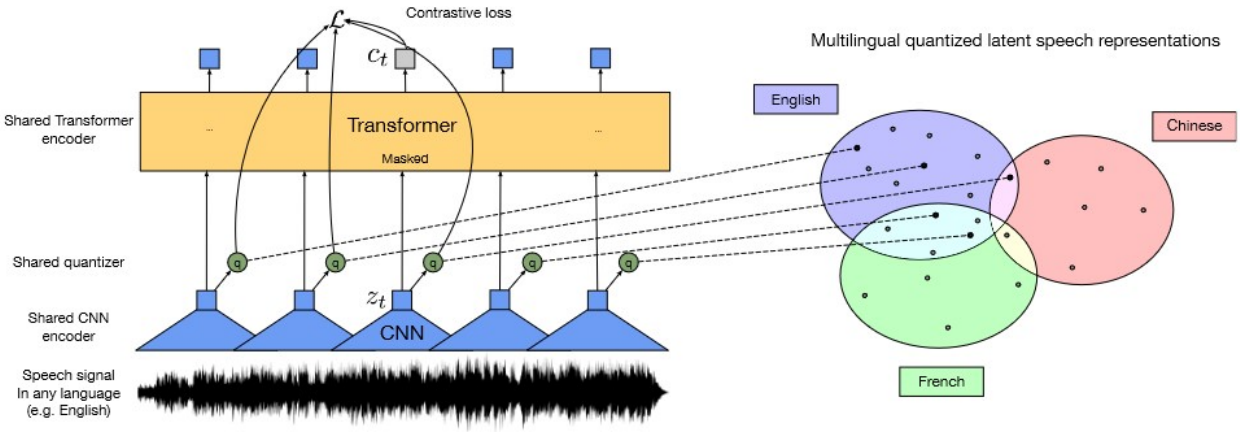

Figure 2.6: Wav2Vec2 XLSR model architecture, figure taken from [24]

The datasets used for pre-training are the CommonVoice dataset with 2k hours, the BABEL dataset with 650 hours of audio, and at last, the Multilingual LibriSpeech(MLS) dataset containing 50k hours of speech. As some of the languages were overrepresented in the complete dataset they had to be penalized in the pre-training to allow for better generalization across the languages. [24]

Our thesis will primarily use this model as the basis for all experiments.

2.4 Metrics

To evaluate the models created in this thesis we require some form of metrics that are comparable and at the same time can give us specific information about the performance. For a better understanding of those metrics, we need to give some context about each one used in this thesis.

2.4.1 Accuracy

Accuracy shows the proportion of correct predictions compared to the total number of predictions with the formal definition seen in equation 2.1. A perfect model that classifies everything correctly has an accuracy of 1.0. [33]

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative} \tag{2.1}$$

This metric has one significant drawback: In multi-class environments, each class has a weight in proportion to its size. Accuracy however does not consider this distribution as well as other indicators which can lead to it hiding strong classification errors. Thus it is generally recommended to use this metric only in binary classifications or switch to an alternative such as the F1-Score.

2.4.2 Precision and Recall

Precision, also called Confidence, is a metric that informs about the proportion of positive identifications or true positives classified by the model which were correct. [33] A perfect model that produces no false positive has a precision of 1.0. The formal definition of precision is given in equation 2.2

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2.2}$$

Recall on the other hand provides insight into the proportion of True Positives that were identified correctly. If the model produces no false negatives the recall is equal to 1.0. [33]

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.3}$$

2.4.3 F1-Score

The F1-Score is a measurement meant to combine precision and recall, which have an inverse relationship, into one single metric. In a binary system, this results in the formal definition found in 2.4

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

The default F1-Score can not be used in a multi-class environment as each class has to have its precision and recall considered. This is possible with either the Macro F1-Score, Weighted F1-Score or Micro F1-Score.

Macro-Averaged F1-Score

The Macro-Averaged F1-Score or Macro F1 in short is used when all classes need to be treated equally and is computed simply by taking the arithmetic mean of each classes Precision and Recall. Using a setup of k classes each class has a recall and precision which are calculated the same as in the previously discussed equations 2.2 and 2.3. Over all k classes a Macro Average Precision and Macro Average Recall is calculated as shown in equations 2.5 and 2.6. [34]

$$MacroAveragePrecision = \frac{\sum_{k=1}^{K} Precision_k}{K} \tag{2.5}$$

$$MacroAverageRecall = \frac{\sum_{k=1}^{K} Recall_k}{K} \tag{2.6}$$

The Macro F1-Score is then calculated by using these averages in the formula 2.4.

$$MacroF1 = 2 * \frac{MacroAveragePrecision * MacroAverageRecall}{MacroAveragePrecision + MacroAverageRecall} \tag{2.7}$$

Because of the averaging, this metric ignores the weights of each class as well. This can be fixed with an additional Metric, the weighted F1-Score.

Weighted-Macro F1-Score

A Weighted F1-Score takes the sample amount n_k of each class and adds these to the precision and recall calculation.

$$WeightedPrecision = \frac{\sum_{k=1}^{K} Precision_k * n_k}{\sum_{k=1}^{K} n_k} \tag{2.8}$$

$$WeightedRecall = \frac{\sum_{k=1}^{K} Recall_k * n_k}{\sum_{k=1}^{K} n_k} \tag{2.9}$$

With this the F1-Score can be calculated again by using the weighted recall and precision in the formula of 2.4.

Micro-Averaged F1-Score

The last variant is the Micro-Average F1-Score. It is best used when there are class imbalances in a multi-class setup. The precision and recall are calculated a bit differently compared to the methods above. Instead of calculating both metrics for each class, they are calculated in one step.

$$MicroPrecision = \frac{\sum_{k=1}^{K} TruePositive_k}{\sum_{k=1}^{K} TruePositive_k + FalsePositive_k} \tag{2.10}$$

$$MicroRecall = \frac{\sum_{k=1}^{K} TruePositive_k}{\sum_{k=1}^{K} TruePositive_k + FalseNegative_k} \tag{2.11}$$

The calculation of the Micro F1-Score is then again the harmonic mean of those to metrics as defined in 2.4. One important fact to consider is that precision will equal recall in the micro-average case which in turn results in an equal F1-Score. Furthermore, this is also the overall accuracy of the model which means that the following formula holds.

$$F1_{Micro} = Precision_{Micro} = Recall_{Micro} = Accuracy \tag{2.12}$$

Chapter 3

Data Collection

3.1 Overview

The dataset we used in this work is from the Swiss Association for Natural Language Processing (SwissNLP) [35] and is currently composed of more than 142'000 samples for a total of about 187 hours. It is a set of sentences of varying sources written in Standard German that have been read out and recorded in the various Swiss German dialects with a duration that can vary between 3 and 10 seconds. Most of these tracks were then voted on and eventually validated. The collection is still in an early stage as their goal is to collect over 2'000 hours worth of data.

The data is supported by a wide range of metadata of which the most relevant include the ID of the user who recorded the speech, the sentence they spoke, their age as well as their gender. To identify the dialect, the canton and the zip code in which the user lives are provided as well. Cantons are always abbreviated and a listing of these can be found in the appendix in Table A.1. The user mean quality gives an idea of the general quality of the audio tracks. The column "clip is valid" indicates if a given sample entry is voted as a correct Swiss German representation of the Standard German sentence.

Following is a deeper analysis of the data in this dataset to evaluate its applicability in training models to distinguish the various dialects. From this analysis, some questions and reflections emerge and some of them will eventually be answered by the experiments. And to conclude the chapter, insight into the criteria that were taken into account to select the test data is given.

3.2 Analysis

3.2.1 Data Partition

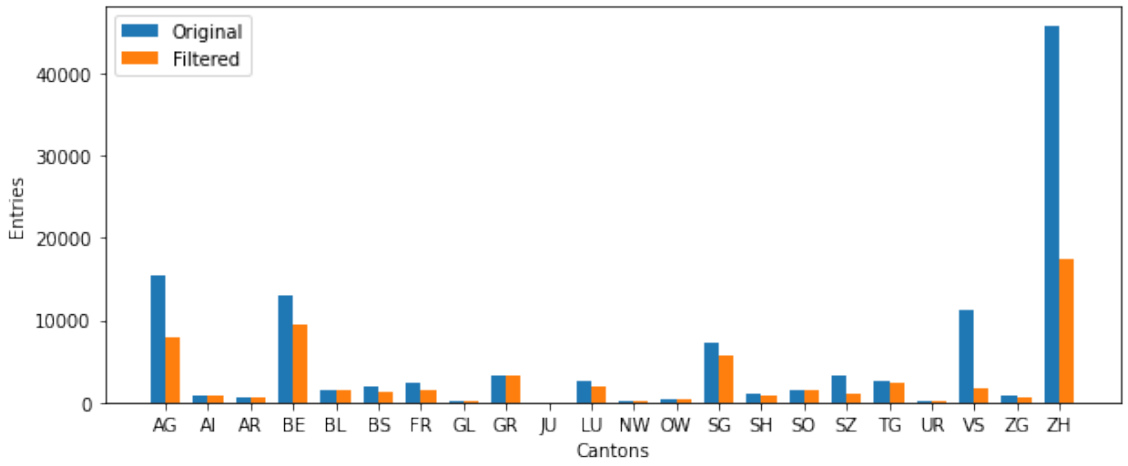

Figure 3.1: Distribution of dataset entries by canton

The partition of the data in the cantons (see Figure 3.1) is not balanced. A small number of cantons have the bulk of the data while others only a few or even none. A deeper analysis, as can be seen in Figure 3.2, shows that several users that hold a sizable portion of the samples of their canton. This could be a problem as the DID model could learn to recognize the user itself rather than distinguishing the dialect. However, it would be interesting to experiment in this direction to see if we can extract dialect features from a single person rather than a group. If so, although unlikely, it would only be necessary to have a small number of users for each canton to be able to distinguish them from each other.

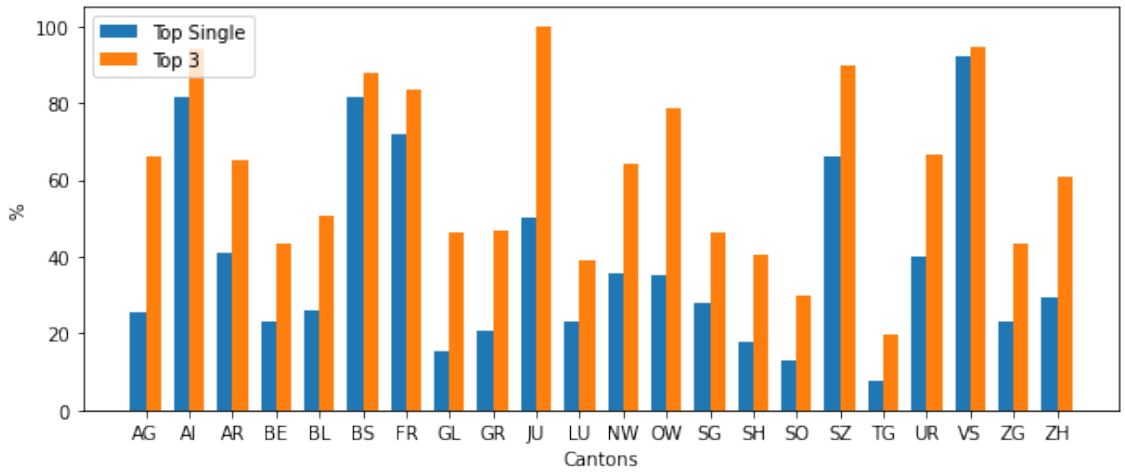

Figure 3.2: Users with the highest number of entries by canton

3.2.2 Gender

Users were able to provide their gender when entering a sample into the speech corpus. A third of them had no gender associated with them. Of those that did, 65'000 were male, 30'000 were female, and 30 samples were of users which indicated themselves as other. Due to the statistical insignificance of the third category 'other', they were not visualized in the following diagrams. The data suggest a certain difficulty in being used in a possible future experiment classifying the gender of an entry without having to discard a large number of samples. Looking at the Figure 3.3 this discrepancy between the genders can be seen among the various cantons.

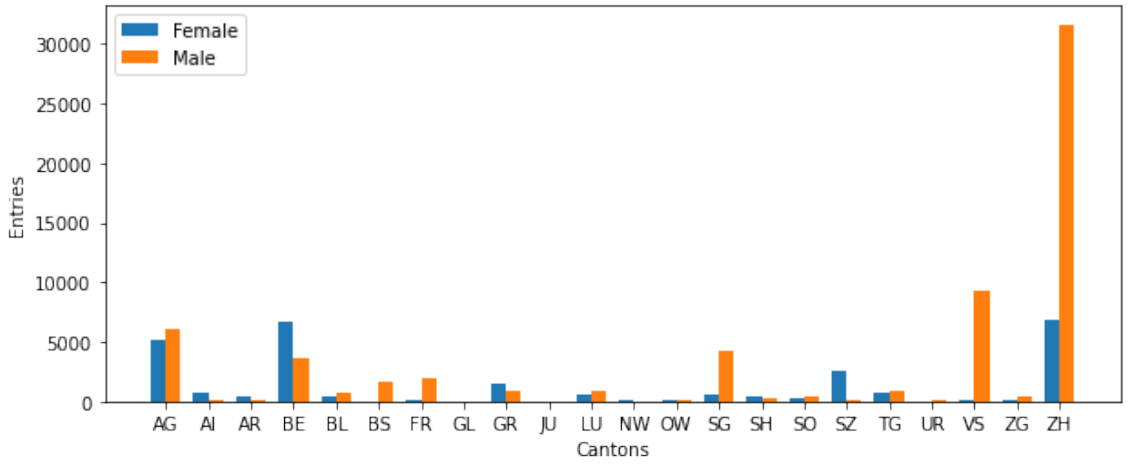

Figure 3.3: Gender distribution by canton

Still, some experiments with female and male categories could be tried with the cantons AG, BE, and ZH as they might have enough data to evaluate the results. This is not to argue that a model is more accurate with one gender than another, but rather to determine whether it is necessary to have a balanced data ratio between the two to train a DID and/or ASR model. It should be noted, however, that theoretically, it is not strictly necessary to use this dataset to answer this question. You should be able to use any other dataset, such as an English or German one, and the result would still be valid.

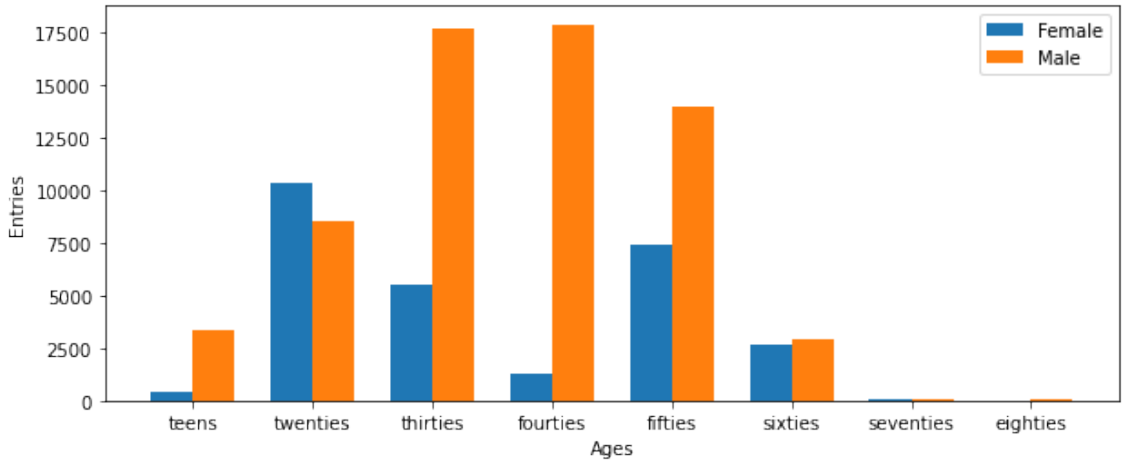

Figure 3.4: Gender distribution by ages

3.2.3 Age

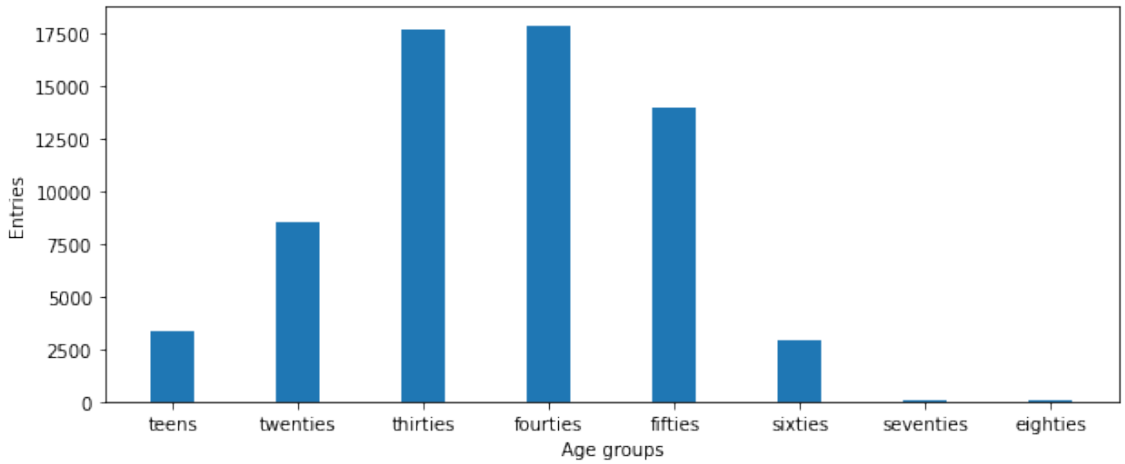

Figure 3.5: Age groups in the dataset

Age is another value that can be considered when training a model that extracts data from voices as these tend to change over the years. The groups most affected are often those at the extremes, in this case at a young age, in the adolescent phase, and partly when approaching seniority. These age groups are the least represented in the dataset (Figure 3.5). There is not enough data left for an accurate assessment when they are assigned to the different cantons. Again, it is not necessary to use this dataset to determine to what degree age groups influence speech recognition models.

3.2.4 Data and Distribution of the Swiss Population

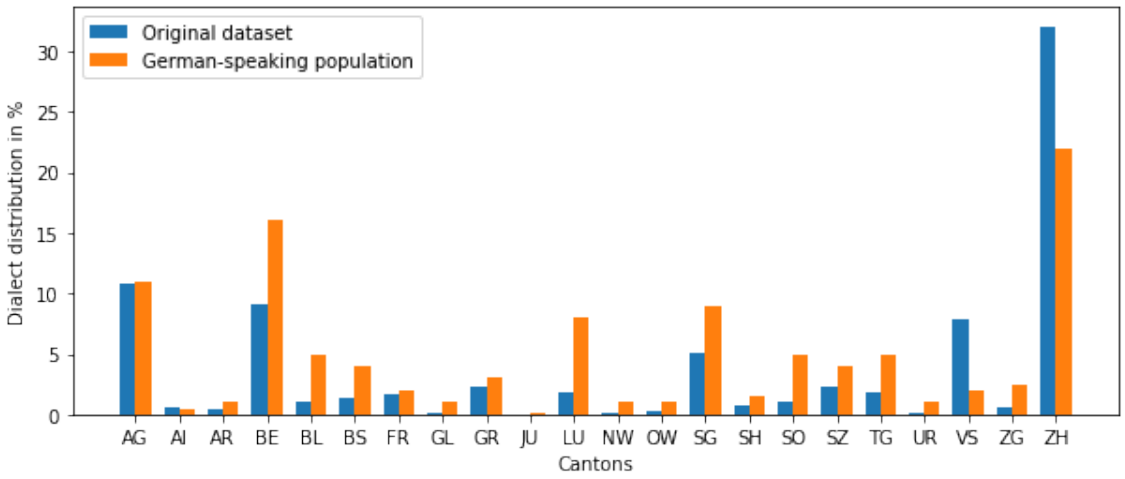

Figure 3.6: Distribution of data in relation to population per canton

If we compare the values of the dataset with the people who speak German in Switzerland (Figure 3.6), we notice a certain degree of imbalance. At the moment, we can argue that it does not represent the distribution of the Swiss population. However, even if they were balanced, this does not necessarily mean anything positive since, by treating each dialect or set of dialects as its own language, it is very likely that better results would be obtained with a fixed number of data. One might rather observe that new data can be expected from cantons that are not very present in the dataset.

3.3 Training Data

Considering the problems in the dataset discussed in chapter 3.2, to select the training data this work will only take part of the big users data into account so that it does not exceed the average amount of data provided per user by an arbitrary factor of 10. To start, data voted invalid and data with a user mean quality of less than 0.7 are discarded. This provides us with a basic filtered training set for the experiments, which will be adapted to the needs of each experiment. Further limitations will generally only be applied concerning the number of samples taken per canton to be able to create evenly distributed tests.

Compared with other work, the best results may be obtained with more data. Consequently, cantons with more data are likely to perform better. If so, it would be interesting to see if and how the cantons could be grouped into larger regions so that they still make sense for the ultimate goal of acting as a sort of switch for the various ASR, making better use of the available data and thus obtaining a more accurate model.

3.4 Test Data

To select the test data, a few points had to be considered. The data should represent the final use of the model, it is assumed, as the final product is a general model, that it will be used with a large variety of users. Consequently, it is desired to have several different users for the tests with few tracks per capita instead of the opposite. It was decided to include tracks with all user qualities as in a real situation audio recording is not always optimal. Regarding the quantity, it was decided to have more or less one thousand data per canton. This way we can have relatively balanced data to perform the different tests. This is applied where the total amount of data allows the filtration, while still keeping reasonably enough data per canton in the training set.

An important point to consider is that the users in the test set must never have been seen during the training. Otherwise, it could be that if the model, especially for experiments where there is not a large variety of users in the cantons, has learned to recognize people and thus the tests are unreliable. This is because in a real case it is difficult for the model to have heard the voices before, or rather it would be out of scope as the model would have to learn which characteristics distinguish different dialects to recognize them, not the people.

Chapter 4

Experimental Setup

4.1 Objective

The objective of this thesis is to test the capabilities of Wav2Vec-XLSR-53 on a Swiss German speech corpus. Specifically, the classification for the individual dialects will be tested. Other tasks such as sex or age classification will, for the time being at least, not be experimented with. These could however be future points of interest. Four distinct granularities will be used to test not only the Wav2Vec model but to also provide an insight into the dialects and their relation to each other. The first test will contain all dialects based on their respective canton which results in 20 different classes. The second and third tests will classify the 10 biggest dialects with the former being a multi-class setup and the latter a binary setup for each combination of the 10 cantons. The last experiment will merge all cantons into four distinct classes based on their regional and linguistic similarities. Based on these four experiments it should be possible to gain deeper knowledge about the performance of Wav2Vec on Swiss German dialects as well as the intrinsic relationships of the dialects themselves.

4.2 Model Selection

When starting this thesis only the base model of Wav2Vec-XLSR-53 with the Mean-Model as a head on top provided by our predecessors [5] were considered to be used. The first test runs with 10 dialects however returned unsatisfactory results with an F1-Score of 10.01% and an accuracy of 14.28%. To verify if these results occurred because of the Mean-Model head or other reasons such as fine-tuning, a switch to the newly released Wav2Vec2ForSequenceClassification [36] linear layer on Hugging Face was tested. The performance was tested again with samples of the 10 biggest dialects. While the results didn't change much, visualized in Figure 4.1, and thus indicating more of a fine-tuning error, a decision was made in favour of the classification head because of the added benefit of it being created and managed by a research engineer at Hugging Face. The model will however still employ the same dimensionality of 1024 for the encoder and projection layer before the mean pooling for classification takes place

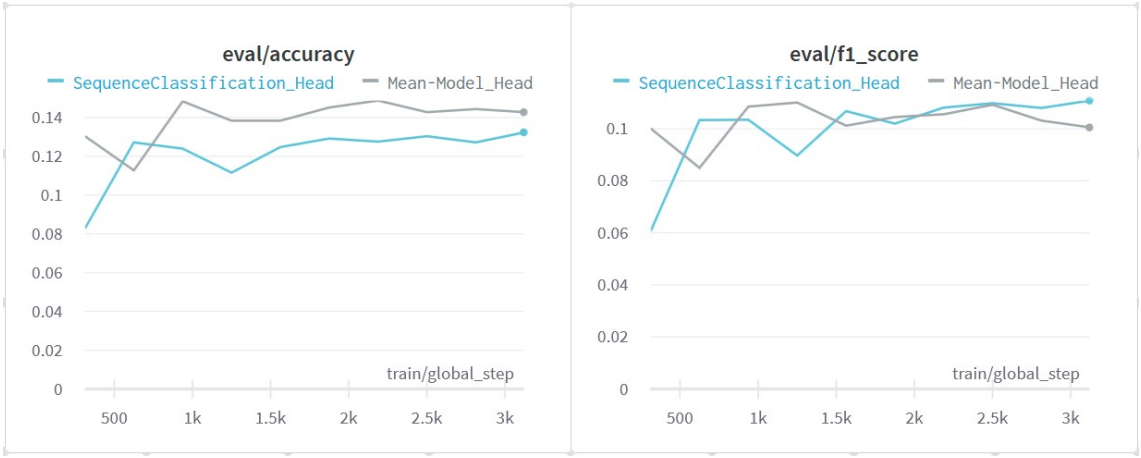

Figure 4.1: Evaluation of pre-trained models

To improve the scores two additional pre-trained Wav2Vec models on HuggingFace were tested. The first was only pre-trained on German data and the second one only used datasets from the Swiss German parliaments of Bern, Zurich, and more provided by the University of Northwestern Switzerland (FHNW) [37]. This time the experiment was kept smaller using only three cantons. The Swiss German model outperformed the base and German model by a clear margin which can be seen in Figure 4.2. Based on these results the decision was made to only use the Swiss German model for the experiments.

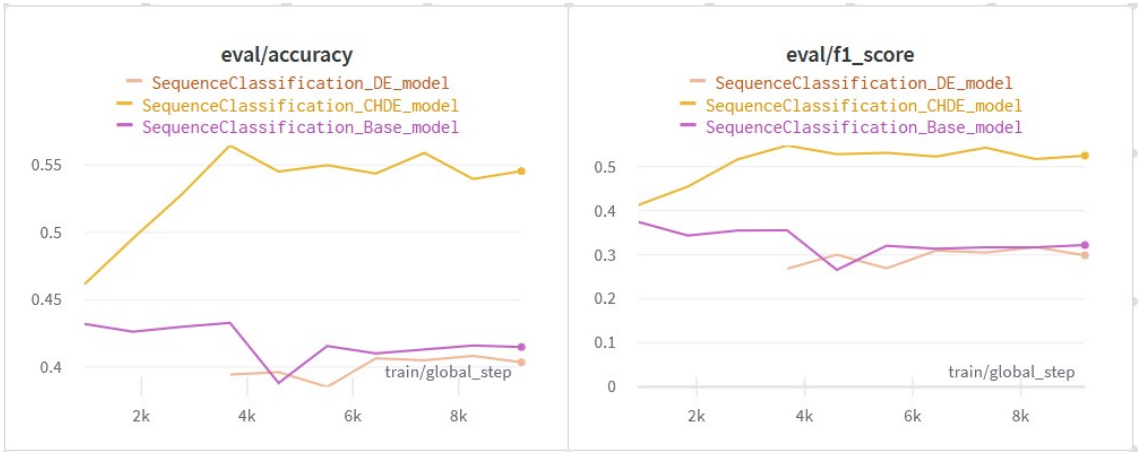

Figure 4.2: Evaluation of pre-trained models

4.3 Corpus

Experiments will exclusively be done with the given Swiss German corpus described in chapter 3. Important to consider is that currently no separation of canton samples will be performed, each canton will be seen as a dialect instead. While the separation of dialects is fluid between the individual canton borders, the separation attempt is beyond the scope of this thesis. We note that there might be some benefit in specifying the correct linguistic borders and thus could be part of a future thesis. A further decision was made to combine the cantons of Basel Stadt and Basel Land on the basis that the area comprising Basel City is very small, the smallest in Switzerland.

4.4 Metrics and Evaluation

The experiments will be measured by the metrics defined in section 2.4, namely the F1-Score. Additional information is provided with the accuracy metric, it will however not be used as a comparison mechanism. Metrics and diagrams will logged by the Weights & Biases evaluation tool. It allows for seamless visualisation, comparison, and tracking of our experiments. [38]

4.5 Training Details

The data is first converted to the WAV format and resampled to 16kHz as this was also the rate on which the Wav2Vec model was pre-trained [23]. To reduce the strain on RAM during training the samples are then stored in h5df files using the h5py python library in a pre-processing step. During training only the required samples are then read into memory again. All models will have the configurations described in Table 4.1 which were tested and verified in an elimination process. Experiments with a lot of data will be trained on two GPUs instead of one which increases the effective batch size to 32.

Parameter	value
Learning rate	$3e^{-6}$
Training epochs	25
Training Batch size	2
Gradient Accumulation steps	8
Warmup steps	200
Save steps	1000
Evaluation steps	500

Table 4.1: Default training parameters for experiments

4.6 Experiments

As described in section 4.1 four different granularities will be tested. The reason for this is the apparent unequal distribution of samples per canton that can be seen in Figure 3.1. Experimenting on these different combinations should give us a deeper understanding of how the individual dialects can be combined or separated to achieve better results in training.

4.6.1 All vs. All

The first experiment will include all data from all 21 Swiss German speaking cantons to see how the model performs on a big unequally distributed multi-class setup. It should further give insight into potential biases from the Swiss German model provided by FHNW [37].

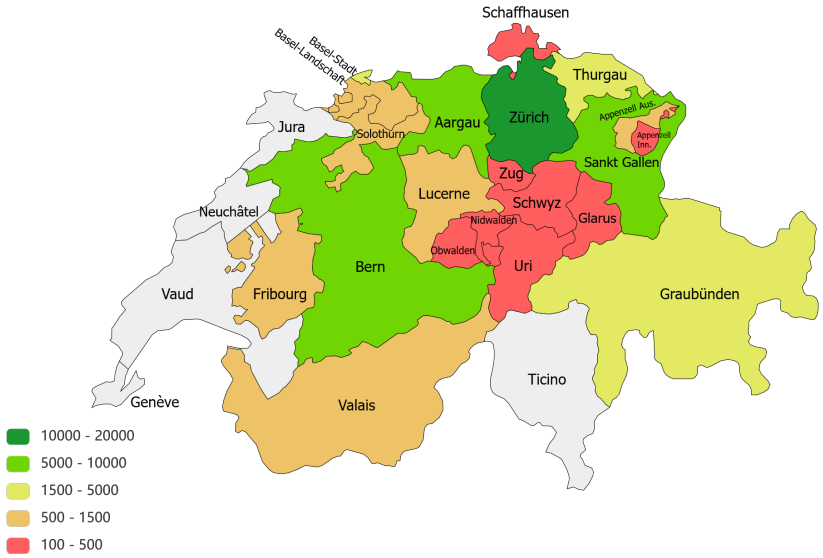

Figure 4.3: Sample distribution per canton after filtering

4.6.2 Big vs. Big

The second experiment tries to test the model's capabilities on an evenly distributed dataset with the 10 biggest cantons that each has up to 1'500 samples. The amount of data was limited by the fact that the smallest canton in this experiment, namely Lucerne, barely has 1'300 samples and thus the decision was made to allow a certain degree of variance as shown in Table 4.2. Cantons used in this experiment are, in alphabetical order, Aargau, Basel which consists of Basel-Stadt and Basel-Land, Bern, Fribourg, Grisons, Lucerne, Saint Gallen, Thurgau, Valais, and Zurich, all of which can be seen in Figure 4.4.

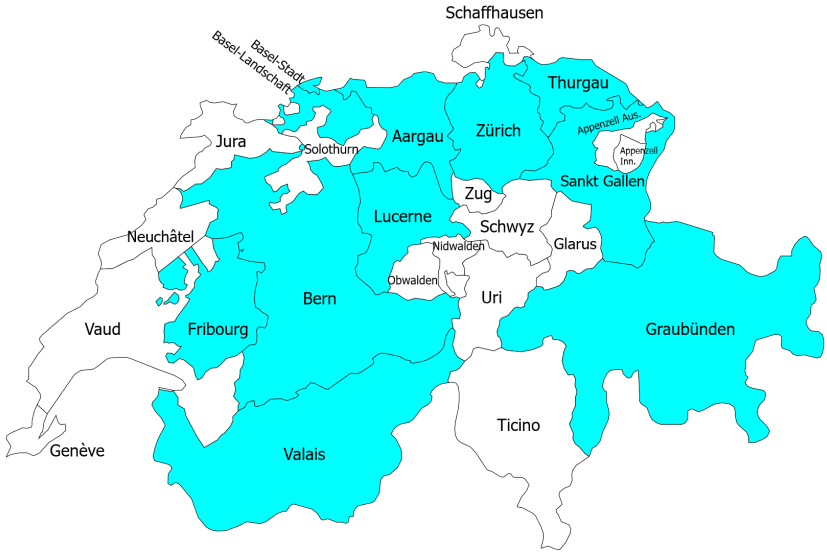

Figure 4.4: 10 biggest dialect classes

Nr.	canton name	abbreviation	sample size of canton	sample size in experiment
1	Aargau	AG	8'006	1'500
2	Basel	BS	2'367	1'500
3	Bern	BE	9'380	1'500
4	Fribourg	FR	1'446	1'446
5	Grisons	GR	2'580	1'500
6	Lucerne	LU	1'280	1'280
7	Saint Gallen	SG	5'223	1'500
8	Thurgau	TG	1'520	1'500
9	Valais	VS	1'420	1'420
10	Zurich	ZH	18'534	1'500

Table 4.2: Training sample distribution 10 biggest cantons

4.6.3 Binary classifications

This third experiment classifies binary combinations of the same 10 cantons as in experiment 4.6.2. By interpreting the results of these experiments we hope to make a more informed decision on how to split the cantons into distinct linguistic groups in experiment 4.6.4. Not all classes have the same amount of samples and as such training was only performed with the maximum amount of available samples of the smaller canton for each canton. Hence the total sample size for each experiment will vary between 3'000 and 8'000 which can be calculated by looking at Table 4.3.

Nr.	max sample amount	cantons
1	1'500	FR, LU, TG, VS
2	2'000	BS
3	2'500	GR
4	4'000+	AG, BE, SG, ZH

Table 4.3: Maximum amount of samples per experiment and canton

4.6.4 Regional Dialects

This fourth and last experiment aims to solve the problem of the regional similarities between cantons which could impair the training. By combining regionally and linguistically similar cantons together a much better score should be able to be achieved. In [39] a connection was able to be made between the individual dialects of Switzerland. Three distinct dialect groups were defined: eastern High-Alemannic German, western High-Alemannic German, and the Highest-Alemannic German. This result should be supported further by the results of the third experiment in section 4.6.3

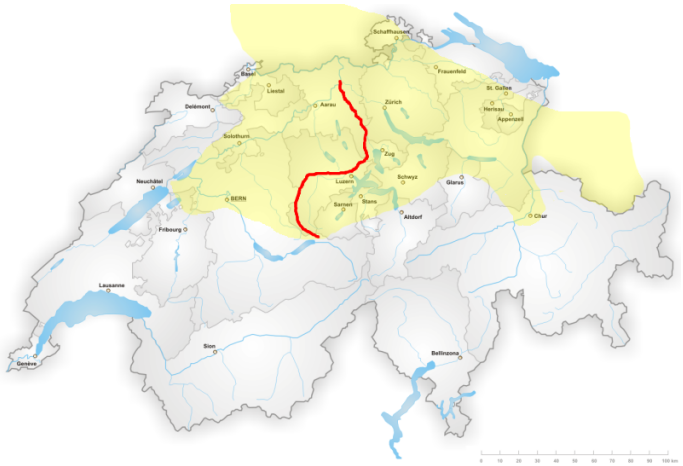

Figure 4.5: Brünig-Napf-Reuss line in red, High-Alemannic area in yellow, figure taken form [40]

A discussion is to be had about the eastern High-Alemannic dialects. The official language area of this dialect group is defined as everything to the east of the Brünig-Napf-Reuss line [41] seen in figure 4.5, to which Zurich and the eastern part of Aargau belong as well. However, the linguistic and also historical relationship of the so-called "Ostschweizer Dialekte" or Eastern Swiss dialects of the cantons Thurgau, Saint Gallen, and others is much closer to each other than to those of Zurich, Aargau et al. [39] Because of this relationship the decision was made to make the "Ostschweizer Dialekte" a separate group, colored blue in figure 4.6, while combining Zurich and the surrounding cantons, to which so-called "Übergangsmundart Kantone" or dialect-transitioning cantons belong to as well, together. For a bet-

ter differentiation between these two groups, we termed the latter as central High-Alemannic, seen in yellow in 4.6. The dialects spoken in the central High-Alemmanic area blur the line between eastern and western High-Alemannic German and make a classification to either west or east difficult and should thus be recognizable by the model as a separate class.

Figure 4.6: Four regional dialect groups

One interesting group are the so called Highest-Alemannic German dialects which are characterized by being spoken by the alpine people of Switzerland and are visualized in red in Figure 4.6. Consisting mostly of cantons with populations high in the Alps like Valais, Uri et al. these dialects are hard to understand even for Swiss people themselves. A separation based not on cantons but on individual villages, towns and cities would be beneficial for this group in a future experiment as not every citizen of these cantons speaks a dialect from Highest-Alemannic.

Nr.	name	abbrev.	cantons	sample size
1	Highest-Alemannic	HA	GL, NW, OW, SZ, UR, VS	2363
2	Western-High-Alemannic	WA	BE, BS, FR, SO	13926
3	Central-High-Alemannic	CA	AG, LU, ZG, ZH	27740
4	Eastern-High-Alemannic	EA	AI, AR, GR, SG, SH, TG	10871

Table 4.4: Regional dialect groups definition

Chapter 5

Results

This chapter provides an analysis of the results gained by the experiments performed in section 4. The metric used for comparisons between the experiments is the macro F1-score as accuracy is not applicable most of the time because of imbalanced test sets. All graphs, diagrams, etc. will reference the macro F1-score unless stated otherwise. Because of time constraints, no repetition for a more stable result of the experiments was able to be made.

5.1 All vs. All

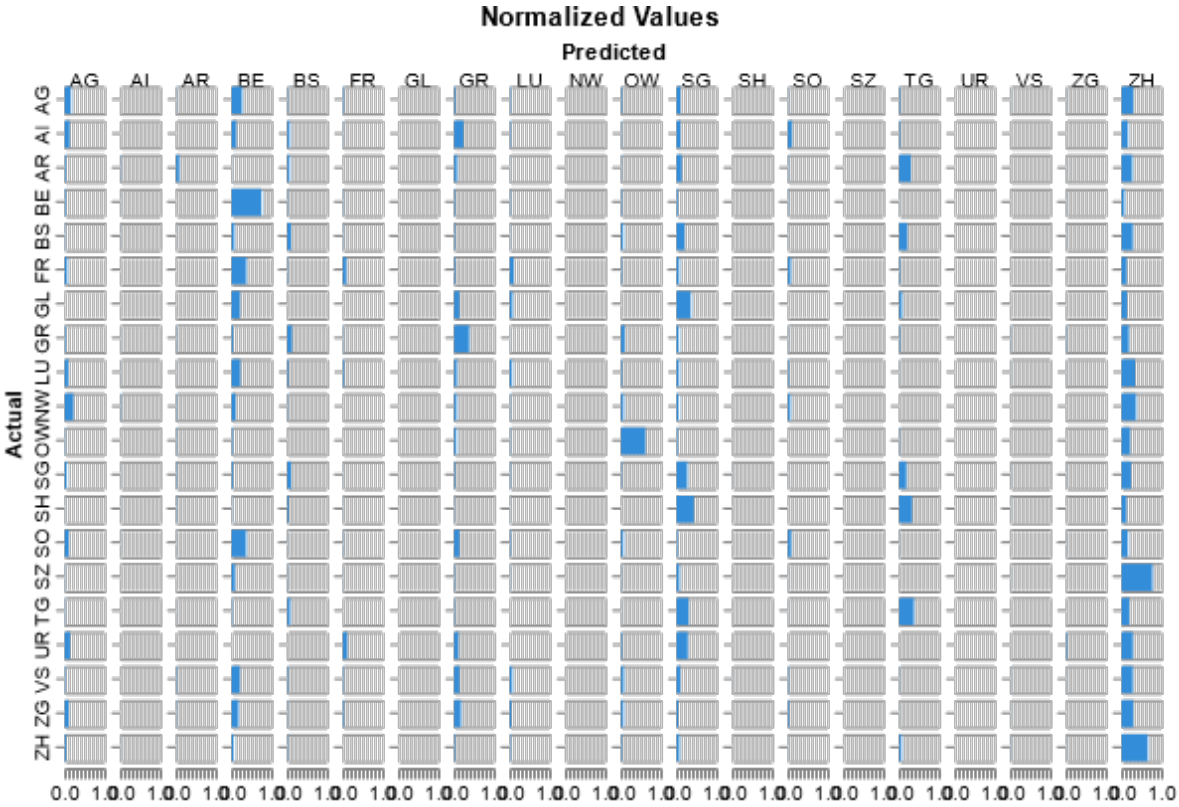

Figure 5.1: All vs. All confusion Matrix

The confusion matrix (Figure 5.1) of the all vs. all experiment shows good results only for a couple of cantons and not for the others. On one hand, we have BE, GR, OW, and TG that seem to perform well, on the other we have several that were not recognised even once, such as AI or GL.

Canton	precision	recall	F1-score
AG	0.25	0.16	0.19
AI	0.00	0.00	0.00
AR	0.25	0.09	0.13
BE	0.34	0.75	0.47
BS	0.15	0.10	0.12
FR	0.24	0.09	0.13
GL	0.00	0.00	0.00
GR	0.31	0.38	0.34
LU	0.20	0.06	0.09
NW	0.00	0.00	0.00
OW	0.11	0.62	0.18
SG	0.18	0.27	0.22
SH	0.00	0.00	0.00
SO	0.28	0.09	0.13
SZ	0.00	0.00	0.00
TG	0.32	0.38	0.34
UR	0.00	0.00	0.00
VS	0.12	0.00	0.01
ZG	0.00	0.00	0.00
ZH	0.18	0.66	0.29

Table 5.1: All vs. All experiment scores of the cantons

The cantons in Table 5.1 with the highest accuracy are BE, TG, and GR. The range 0.30-0.20 follows with SO, AG, AR, FR, and LU. Intuitively it might be said that OW has high precision, but it must be taken into account that the confusion matrix shows normalised values. As far as recall is concerned, we find good results for BE, ZH, OW. As for GR, TG, SG, the results are discrete. To have a more balanced evaluation between precision and recall, we pass to the F1-score and remark that it gives more weight to the lower of the 2 values. Here we have acceptable values for BE, GR, TG, and ZH.

F1-score	accuracy
13.24%	23.70%

Table 5.2: All vs. All experiment score

The accuracy of this model (Table 5.2) is 23.70% which, however, being an experiment with unbalanced evaluation data, is to be taken with caution. While the macro F1-score is 13.24%, on the other hand, a better indicator for this case might be the weighted F1-score with 0.18.

5.1.1 Conclusions

The model is generally unconfident and difficult to use to identify cantons. One result where more would have been expected due to the number of audio tracks available is Zurich (ZH) which has an F1-score of 0.29. The amount of available data may have an influence only to a certain degree.
Having taken all available filtered data of all present cantons, it was not possible to have the same number of data for all cantons. Except for VS and OW, all cantons with an F1-score close to 0 are all those with less than 500 audio tracks.
Several cantons have low confidence, but looking at the confusion matrix it can be seen it happens that incorrectly recognised tracks are part of neighbouring cantons or dialect categories. Considering the traces classified by the model as Bern (BE), we also find several traces from FR and SO which are part of the same regional group, and also LU and AG which are neighbouring cantons. This suggests that grouping by region may be beneficial for recognising dialect types.

5.2 Big vs. Big

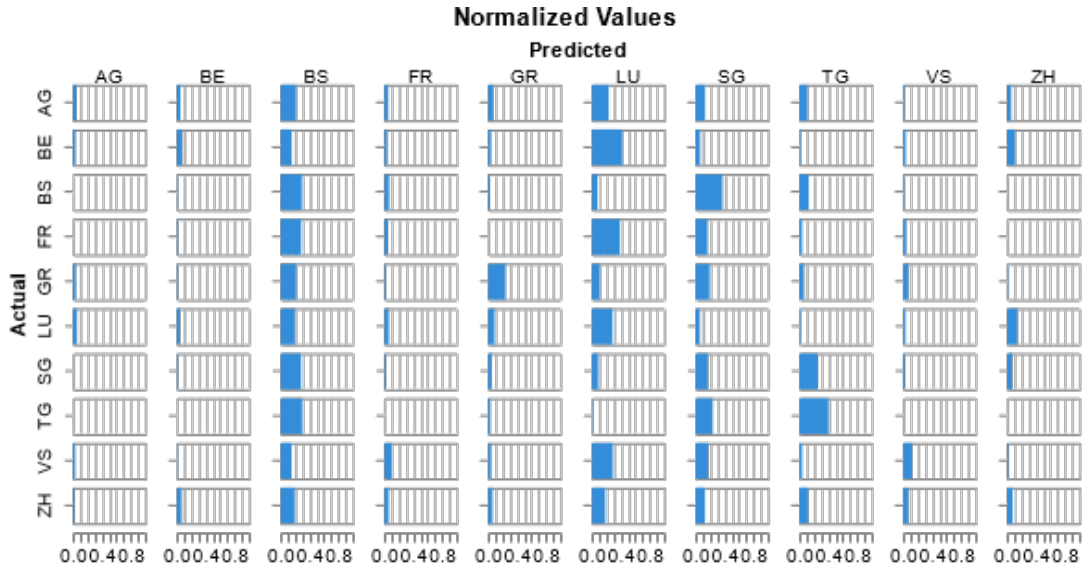

Figure 5.2: Big vs. Big confusion Matrix

Looking at the result achieved by training the model with the 10 largest cantons balanced with about 1500 elements, we find a certain inconsistency in the identification of cantons. Looking at the confusion matrix, the best recognised cantons are TG and GR. The model visibly struggles to recognise BS, LU, and SG. The remaining ones are hardly classified.
From the Table 5.3 we see that among the values higher than 0.2 of accuracy, besides TG and GR, there are only BE, VS and AG. However, they all remain below 0.35 which makes the model unconfident. TG is recognised by the model at 40%, followed by BS with a recall of 0.3, LU with 0.28, GR with 0.24, and the others below 0.2. Only TG and GR have an acceptable f1-score with 0.37 and 0.27 respectively.

Canton	precision	recall	F1-score
AG	0.21	0.05	0.08
BE	0.25	0.08	0.12
BS	0.12	0.30	0.17
FR	0.06	0.05	0.05
GR	0.31	0.24	0.27
LU	0.16	0.28	0.20
SG	0.12	0.17	0.14
TG	0.34	0.40	0.37
VS	0.25	0.13	0.17
ZH	0.16	0.08	0.10

Table 5.3: Big vs. Big experiment scores of the cantons

F1-score	accuracy
16.87%	18.09%

Table 5.4: Big vs. Big experiment score

While the accuracy value cannot be taken into account for this type of experiment, the macro F1-score is quite low at 16.87% (see Table 5.4) the macro f1-score is quite low at 16.87% along with a similar weighted F1-score of 0.17.

5.2.1 Conclusions

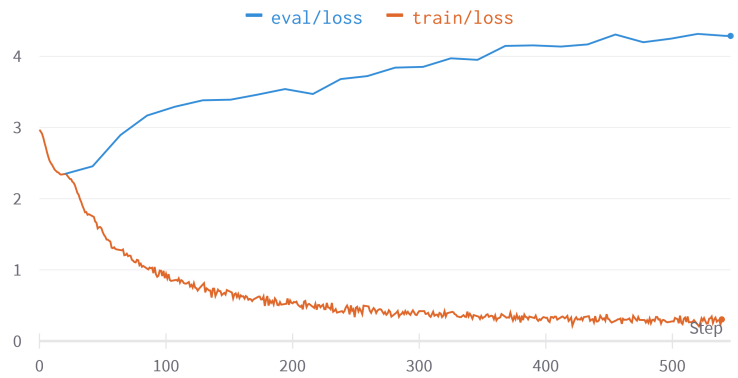

Figure 5.3: Training and evaluation loss function for Big vs. Big experiment

This experiment demonstrated a certain difficulty in distinguishing between the various cantons. In part, this could be due to the similarity of dialects, if not the same, spoken in multiple cantons. It is worth mentioning the graph of the loss function (Figure 5.3) as it shows an upward curve in the evaluation data part. This could be due to several causes as the test data are not representative of the experiment. Having the relatively correct training loss curve, one could interpret

this discrepancy as an overfit model that consequently struggles to represent the test data correctly. Therefore it remains feasible that to improve the result, further fine-tuning work should be carried out and the data used for training and testing should be more closely investigated.

5.3 One vs. One

The One vs. One experiment was conducted to enable us to give a more informed decision on how a split of cantons into four distinct groups for the regional dialects experiment should be performed as well as to provide us more insight into the similarities of the dialects themselves. A total of 45 individual runs for each possible combination of the 10 largest cantons were made. All of these results can be viewed in the appendix in Table A.2. This section will go into more detail about the outliners which either performed very well, visualized in Table 5.5, or very poorly as shown in Table 5.6.

Nr.	canton A	canton B	F1-score	accuracy
1	SG	BE	89.63%	89.63%
2	LU	TG	87.91%	87.92%
3	TG	VS	86.55%	87.19%
4	BE	TG	85.58%	85.77%
5	FR	TG	84.31%	86.87%

Table 5.5: Top 5 results of the One vs. One

What is most interesting in Table 5.5 is the occurrence of Thurgau (TG), an eastern High-Alemannic dialect, in four of the total five best runs. Each of these four runs had a western High-Alemannic dialect or at least a canton that is partially occupied by western dialect-speaking citizens in the case of Lucerne (LU), as its opposite. The top result with an F1-score of 89.63% and an accuracy of 89.63% concerning the comparison between Saint Gallen (SG) and Bern (BE) is also a classification of an eastern and a western dialect respectively. This indicates a general distinguishability for eastern and western High-Alemannic dialects which was expected.

Nr.	canton A	canton B	F1-score	accuracy
41	AG	ZH	47.77%	53.45%
42	LU	ZH	44.40%	50.15%
43	BE	FR	43.74%	64.67%
44	TG	ZH	41.50%	54.84%
45	AG	LU	40.07%	52.49%

Table 5.6: Bottom 5 results of the One vs. One experiment

The Table 5.6 shows multiple cantons of the in section 4.6.4 proposed central High-Alemannic group as the worst-performing combinations. This indicates a general

similarity between these cantons and further supports the theoretical grouping defined in section 4.6.4. An interesting aspect is the occurrence of Thurgau (TG) which is close to the canton of Zurich and thus results in a bad score. Apart from this combination however all results in comparisons are made up of cantons in the same regional group. For a definitive answer on the applicability of these proposed regions, it should be analyzed and experimented upon again in future works to rule out possible errors in fine-tuning or the usage of bad samples.

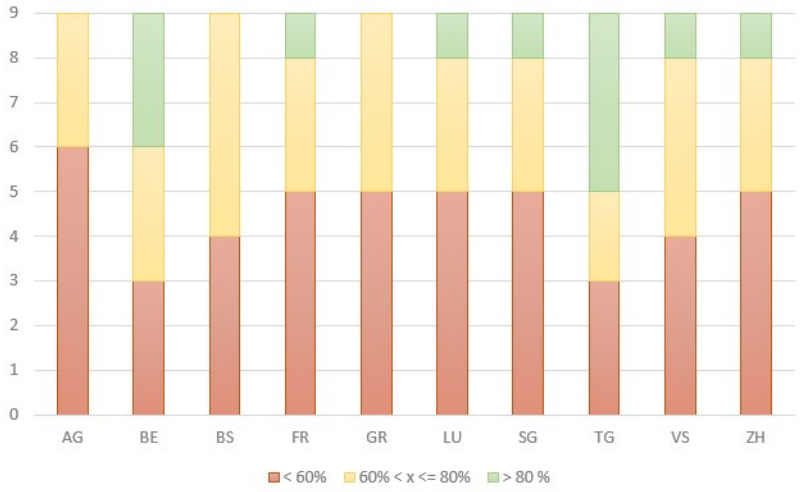

Figure 5.4: Distribution of F1-Scores per canton

Figure 5.4 gives a more comprehensive overview of the results by ordering them into three categories based on the F1-score per canton. Green are results with an F1-Score over 80%, yellow between 60% and 80%, and red everything under 60%. In Table 5.7 the average F1-Score and accuracy are shown.

F1-score	accuracy
62.74%	66.42%

Table 5.7: Average scores in the One vs. One experiment

5.3.1 Conclusions

This experiment showed that eastern and western dialects are distinguishable from each other. It enforces our assumption of Wav2Vec being able to differentiate Swiss German dialects if enough data is available for training. It mostly reinforces our theoretical defined boundaries of the regional dialects as well. However, because none of the experiments were able to be repeated these findings should be viewed with a certain degree of caution. Future works should rerun these experiments multiple times and give a more detailed answer to this task.

5.4 Regional dialects

The performance of the regional dialect experiment was highly anticipated as it should answer some of our theoretical assumptions about the linguistic similarities of the Swiss German dialects. The experiment yielded an accuracy of 52.89%, a macro F1-score of 45.95%, and a weighted F1-score of 0.5. The detailed results are visualized in the confusion matrix 5.5 and the Table 5.8.

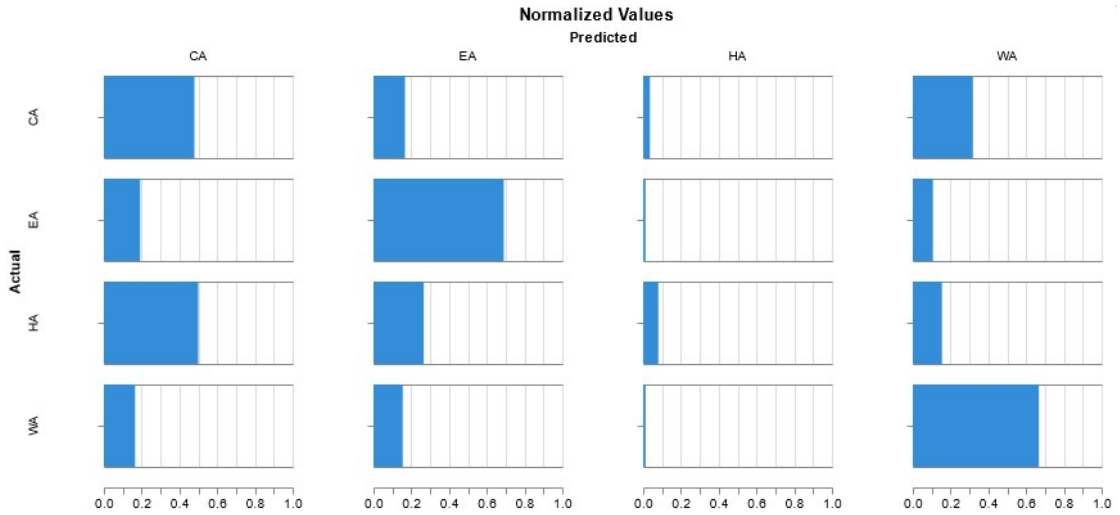

Figure 5.5: Confusion matrix regional dialects

Canton	precision	recall	F1-score
EA	0.62	0.69	0.65
CA	0.44	0.48	0.46
WA	0.53	0.67	0.59
HA	0.42	0.08	0.13

Table 5.8: Precision, Recall and F1-score of the different regions

A more in-depth analysis will be given here to each of the four regions. The model scored the best results with the eastern High-Alemannic region (EA) region with an F1-Score of 0.65. It mainly struggled with identifying test samples as part of the central High-Alemannic (CA) group which was to be expected based on the proximity of the two regions. This result reinforces our decision of grouping the "Ostschweizer Dialekte" together instead of using a different combination that would have included the whole of Zurich. Highest-Alemannic (HA) was never even classified with EA samples.

The second-best result with an F1-score of 0.59 is that of the western High-Alemannic (WA) group which had an equal proportion of its false positives in both CA and EA which is surprising considering that linguistic similarities should be low for EA and WA dialects. We expected the results to look more like those of the EA class

which had sometimes difficulties keeping CA and itself apart. One characteristic that continued here as well was the extremely low misclassifications of samples as part of the HA region.

The model had some difficulties with the CA group in the third place as it liked to misclassify the test samples as WA. It should be noted that the CA group is a mix between eastern and western dialects thus some error was prone to happen. This group should be analyzed again based on a zip code level and not on a canton level to allow for some better grouping of samples.

By far the worst results were created when classifying HA samples. It consistently failed at recognizing its class and instead assigned them to the CA group which was not expected. Though similarities exist based on the regional closeness and the fact that not all cantons in the HA region exclusively speak a dialect of that family, those similarities should not be pronounced enough to amount to such a strong misclassification. An additional fact to consider is the low amount of resources available for this region. Only around 3'000 samples were able to be used in the experiment while the other three regions each had more than 10'000. This experiment could thus be run again in an evenly distributed fashion to see if the classifications of HA dialects could benefit in a fair setup.

5.4.1 Conclusions

While the results were not entirely satisfying they did show promising prospects for further experiments using this regional setup. By defining the lines in a more detailed manner and not on a canton level we believe even better results should be possible. The issue with the HA region must be investigated closely by either using an evenly distributed setup or by having more samples available. Considering that data collection is difficult in that region a decision has to be made in the future if the HA region should be ignored for the time being while further developments are made for the eastern, central, and western dialects.

Chapter 6

Discussion and Outlook

This thesis aimed at testing the capabilities of Wav2Vec on a newly released low resource speech corpus of Swiss German. While low resource tasks have been solved before, however few used spoken Swiss German or Wav2Vec as its basis. By combining different experiments we tried to gain more insight into the potential of Wav2Vec-XLSR-53 as a DID model, understanding the relationships of the individual Swiss German dialects as well as performing a first analysis of the new speech corpus provided by SwissNLP.

The in-depth analysis of the speech corpus gave some insight into the distribution of the data. It showcased that a small number of cantons hold a significant share of the data which is not an advantage to the task at hand. When gender and age of users were considered it was discovered that there is an imbalance between male and female entries. Currently, no verifications have been done if this imbalance influences the findings of this thesis. There were also concerns with some users being overrepresented in their respective canton which lead to a lot of samples being filtered out for training.

An important fact needs to be addressed first before taking the experimental findings at face value. Because of the time constraints posed upon this thesis, none of the four experiments were able to be repeated. This may mean that there is a lack of reproducibility and a certain degree of uncertainty in the results. Future work should build upon this thesis and verify the results described here.

While not all experiments returned satisfying results, each one had an interesting characteristic. In the first experiment, where all cantons were pitched against each other, only three cantons had an F1-score above 0.3. Thurgau (TG) stood out among all three canton-based experiments. In all of them it performed relatively well, even being the best performing in the One vs. One experiment. This indicates that the collected data of Thurgau could either be of high quality or the model can distinguish it uniquely well from other dialects. The region it was grouped into in experiment four, the eastern High-Alemannic group, also performed the best out of the rest.

The result of the regional experiment, which was the most interesting one to us, mostly confirmed our theoretical linguistic borders of the Swiss German dialects. A major exception of that was the Highest-Alemannic group of which we do not currently know the exact reason as to why it performed so poorly. Some obvious assumptions like the unevenly distributed dataset or the fluid boundaries of the dialects within the cantons of the group can be made, but have yet to be verified. The other three regions however returned mostly satisfying results with the central group having the most difficult task of being a transition zone of eastern to western dialects and thus lowering the score by misclassifying some samples. This generally displays that classification of Swiss German dialects into regions is possible.

In the one vs. one experiment, a lot of the scores were in an expected range considering the geographic proximity of most of the cantons. Tough a few outliners need verification as to why they performed either exceptionally well or poorly, it generally was supportive of the theoretical boundaries defined in the regional experiment

We showed that the results gained by the experiments are implying that Wav2Vec is in general capable of classifying Swiss German dialects. While certain fine-tunings and verifications have still to be made, it is nonetheless a promising start. With the release of the new Wav2Vec-XLS-R model, which further increased the performance of the XLSR-53 model used in this thesis, we are confident that a dialect-based speech recognition system is possible for Swiss German. In a possible continuation of the work, it is envisaged to train two separate systems for the different dialects, a DID model and an STT model. The DID system detects the dialect spoken and selects the correct STT system which translates the classified audio track into text. In this case, further studies would have to be made to find out what the most meaningful way of grouping the cantons is, taking into account different factors such as the grammar of the various dialects. A on similar dialects trained model is most likely more successful at translating speech into text than one which was trained with samples that do not correlate with each other. Experiments will have to assess how well the combination of these DID and STT models perform and see if better results are obtained compared to other research.

Bibliography

[1] FederalStatisticalOffice, "Main languages of the permanent resident population, 1970-2019." [Online]. Available: https://www.bfs.admin.ch/bfs/en/home/statistics/population/languages-religions/languages.assetdetail.15384182.html

[2] M. Weibel and M. Peter, "Compiling a Large Swiss German Dialect Corpus," in *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. Schweizerische Idiotikon, Jun. 2020. [Online]. Available: https://aclanthology.org/2021.wanlp-1.28

[3] M. Szmigiera, "The most spoken languages worldwide in 2021," https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

[4] M. Plüss, L. Neukom, and M. Vogel, "Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus," *CoRR*, vol. abs/2010.02810, 2020. [Online]. Available: https://arxiv.org/abs/2010.02810

[5] P. Fivian and D. Reiser, "Speech classification using wav2vec 2.0," [Online; Accessed 16.12.2021]. [Online]. Available: https://www.zhaw.ch/storage/engineering/institute-zentren/cai/BA21_Speech_Classification_Reiser_Fivian.pdf

[6] A. Tjandra, D. G. Choudhury, F. Zhang, K. Singh, A. Baevski, A. Sela, Y. Saraf, and M. Auli, "Improved language identification through cross-lingual self-supervised learning," *CoRR*, vol. abs/2107.04082, 2021. [Online]. Available: https://arxiv.org/abs/2107.04082

[7] O. Zaidan and C. Callison-Burch, "Arabic Dialect Identification," *Computational Linguistics*, vol. 40, no. 1, pp. 171–202, Mar 2014. [Online]. Available: https://doi.org/10.1162/COLI_a_00169

[8] M. Zampieri, S. Malmasi, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, Y. Scherrer, and N. Aepli, "Findings of the VarDial evaluation campaign 2017," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics, Apr. 2017, pp. 1–15. [Online]. Available: https://aclanthology.org/W17-1201

[9] R. Ionescu and A. Butnaru, "Learning to identify arabic and german dialects using multiple kernels," in *Proceedings of the Fourth Workshop on*

NLP for Similar Languages, Varieties and Dialects (VarDial). Association for Computational Linguistics, Apr. 2017, pp. 200–209. [Online]. Available: https://aclanthology.org/W17-1225"

[10] T. Jauhiainen, H. Jauhiainen, and K. Lindén, "HeLI-based experiments in Swiss German dialect identification," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. Association for Computational Linguistics, Aug. 2018, pp. 254–262. [Online]. Available: https://aclanthology.org/W18-3929

[11] F. Weninger, Y. Sun, J. Park, D. Willett, and P. Zhan, "Deep learning based mandarin accent identification for accent robust asr," in *Proc. Interspeech 2019*, ser. Tech. Rep, Sep. 2019.

[12] R. Huang and J. Hansen, "Gaussian mixture selection and data selection for unsupervised spanish dialect classification," in *Interspeech 2016*, Dec. 2006.

[13] M. Abdul-Mageed, C. Zhang, A. Elmadany, H. Bouamor, and N. Habash, "NADI 2021: The second nuanced Arabic dialect identification shared task," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Apr. 2021, pp. 244–259. [Online]. Available: https://aclanthology.org/2021.wanlp-1.28

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, Oct. 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] M. Abdul-Mageed, A. Elmadany, and E. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 7088–7105. [Online]. Available: https://aclanthology.org/2021.acl-long.551

[16] B.-H. Juang and L. Rabiner, "Speech recognition, automatic: History," in *Encyclopedia of Language & Linguistics (Second Edition)*, second edition ed., K. Brown, Ed. Oxford: Elsevier, 2006, pp. 806–819. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B0080448542009068

[17] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[18] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[19] P. Ivić and D. Crystal, "Dialect," Encyclopedia Britannica https://www.britannica.com/topic/dialect, [Online; Accessed 15.12.2021].

[20] A. Etman and A. A. L. Beex, "Language and dialect identification: A survey," in *2015 SAI Intelligent Systems Conference (IntelliSys)*, 2015, pp. 220–231.

[21] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *CoRR*, vol. abs/1904.05862, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1904.05862

[22] A. Baevski, S. Schneider, R. Collobert, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *Facebook AI*, vol. Tech. Rep., Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.05453

[23] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Facebook AI*, vol. Tech. Rep., Oct. 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[24] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," *Facebook AI*, vol. Tech. Rep., Dec. 2020. [Online]. Available: https://arxiv.org/abs/2006.13979

[25] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, Y. von Platen, P.and Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," *Facebook AI*, vol. Tech. Rep., Nov. 2021.

[26] P. Le-Khac, G. Healy, and A. Smeaton, "Contrastive representation learning: A framework and review," *CoRR*, vol. abs/2010.05113, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.05113

[27] EncyclopediaBritannica, "phoneme," https://www.britannica.com/topic/phoneme, [Online; Accessed 10.12.2021].

[28] W. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *CoRR*, vol. abs/1704.04222, Sep. 2017. [Online]. Available: http://arxiv.org/abs/1704.04222

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, Jun. 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[30] R. Kulshrestha, "Transformers," https://towardsdatascience.com/transformers-89034557de14, [Online; Accessed 10.12.2021].

[31] R. Futrzynski, "Getting meaning from text: self-attention step-by-step video," https://peltarion.com/blog/data-science/self-attention-video, [Online; Accessed 11.12.2021].

[32] L. Sus, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," https://neurosys.com/wav2vec-2-0-framework/, [Online; Accessed 11.12.2021].

[33] D. M. W. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *CoRR*, vol. abs/2010.16061, 2008. [Online]. Available: https://arxiv.org/abs/2010.16061

[34] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *CoRR*, vol. abs/2010.16061, 2020. [Online]. Available: https://arxiv.org/abs/2008.05756

[35] SwissNLP, "Die Schweizer Dialektsammlung," https://swissnlp.org/die-schweizer-dialektsammlung/, [Online; Accessed 12.12.2021].

[36] Huggingface, "Wav2vec for sequence classification," https://huggingface.co/docs/transformers/model_doc/wav2vec2#transformers.Wav2Vec2ForSequenceClassification, [Online; Accessed 16.12.2021].

[37] Y. Kaufmann, "wav2vec2-large-xlsr-53-swiss-german on Swiss German fine-tuned model," https://huggingface.co/Yves/wav2vec2-large-xlsr-53-swiss-german, [Online; Accessed 12.12.2021].

[38] Weights&Biases, "Developer tools for ML," https://wandb.ai/site, [Online; Accessed 16.12.2021].

[39] R. Hotzenköcherle, *Die Sprachlandschaften der deutschen Schweiz*, ser. Reihe Sprachlandschaft Bd. 1. Aarau: Sauerländer, 1984.

[40] Dbachmann, "Brunig-napf-reuss-linie," https://commons.wikimedia.org/wiki/File:Brunig-Napf-Reuss-Linie.png, [Online; Accessed 12.12.2021].

[41] R. Weiss, "Die Brünig-Napf-Reuss-Linie als Kulturgrenze zwischen Ost- und Westschweiz auf volkskundlichen Karten," *Geographica Helvetica*, vol. 2, no. 3, pp. 153–175, 1947. [Online]. Available: https://gh.copernicus.org/articles/2/153/1947/

List of Figures

List of Tables

Appendix A

Experiment Details

canton	abbreviation
Aargau	AG
Appenzell Innerrhoden	AI
Appenzell Ausserrhoden	AR
Bern	BE
Basel Land	BL
Basel Stadt	BS
Fribourg	FR
Glarus	GL
Graubünden / Grisons	AG
Jura	JU
Luzern	LU
Nidwalden	NW
Obwalden	OW
Sankt Gallen / Saint Gallen	SG
Schaffhausen	SH
Solothurn	SO
Schwyz	SW
Thurgau	TG
Uri	UR
Wallis / Valais	VS
Zug	ZG
Zürich	ZH

Table A.1: Swiss German canton abbreviations

Nr.	canton name A	canton name B	samples for each canton	Macro F1	Accuracy
1	SG	BE	4000	89.63 %	89.63 %
2	LU	TG	1500	87.91 %	87.92 %
3	TG	VS	1500	86.55 %	87.19 %
4	BE	TG	1500	85.58 %	85.77 %
5	FR	TG	1500	84.31 %	86.87 %
6	BE	ZH	4000	82.28 %	82.37 %
7	BS	BE	2000	75.43 %	77.06 %
8	SG	FR	1500	74.49 %	79.29 %
9	GR	TG	1500	74.21 %	74.24 %
10	BS	GR	2000	72.00 %	72.00 %
11	AG	VS	1500	70.12 %	70.93 %
12	BS	LU	1500	68.59 %	69.06 %
13	AG	GR	2500	67.53 %	67.53 %
14	BS	VS	1500	67.47 %	70.61 %
15	BE	GR	2500	67.28 %	72.58 %
16	LU	VS	1500	65.52 %	69.77 %
17	SG	LU	1500	64.55 %	66.41 %
18	BE	VS	1500	64.42 %	65.25 %
19	FR	GR	1500	64.23 %	66.09 %
20	FR	VS	1500	63.97 %	64.13 %
21	BS	TG	1500	62.19 %	62.83 %
22	AG	FR	1500	61.48 %	61.59 %
23	VS	ZH	1500	61.14 %	61.49 %
24	SG	ZH	4000	60.69 %	63.58 %
25	FR	ZH	1500	60.31 %	61.17 %
26	SG	GR	2500	58.79 %	62.37 %
27	BE	AG	4000	58.65 %	61.10 %
28	SG	AG	4000	58.53 %	58.73 %
29	SG	BS	2000	58.46 %	60.89 %
30	GR	LU	1500	58.10 %	58.99 %
31	FR	LU	1500	56.63 %	64.02 %
32	SG	VS	1500	56.37 %	65.61 %
33	BS	AG	2000	54.92 %	64.19 %
34	BS	FR	1500	54.03 %	56.52 %
35	AG	TG	1500	53.17 %	60.66 %
36	SG	TG	1500	53.03 %	56.07 %
37	BS	ZH	2500	52.20 %	63.98 %
38	BE	LU	1500	51.74 %	53.91 %
39	GR	VS	1500	49.78 %	49.78 %
40	GR	ZH	2500	48.88 %	62.13 %
41	AG	ZH	4000	47.77 %	53.45 %
42	LU	ZH	1500	44.40 %	50.15 %
43	BE	FR	1500	43.74 %	64.67 %
44	TG	ZH	1500	41.50 %	54.84 %
45	AG	LU	1500	40.07 %	52.49 %

Table A.2: Training results binary setup

Appendix B

User manual

In the project repository on GitHub you can find the user manual at the following link:
https://github.zhaw.ch/Swiss-German-Dialects-Recognition/w2v-ch-de-recognition/blob/master/references/Environment%20Setup.md

Appendix C

Code

The code developed for this is thesis is available on github.zhaw.ch:
https://github.zhaw.ch/Swiss-German-Dialects-Recognition/w2v-ch-de-recognition