Automatic Identification of Swiss German Dialects Via Speech-to-phoneme Transcriptions

Laura Bolliger and **Safiyya Waldburger** and **Mark Cieliebak** and **Jasmina Bogojeska**
bollilau@students.zhaw.ch

Abstract

Within the domain of multilingual speech recognition systems, differentiating between languages and dialects is crucial. This task is particularly challenging for low-resource languages like Swiss German, which is spoken by comparatively few people. Our project explored various approaches for automatically identifying Swiss German dialects from audio data. The objective was to classify speech samples into one of seven dialect regions.

In a first attempt, the pretrained speech recognition system Whisper was fine-tuned on subsets from two recent Swiss German corpora SDS-200 and STT4SG-350. The number of speakers as well as the number of samples per dialect region were varied, and it was found that a larger number of speakers and a balanced number of samples per dialect and speaker are favorable. Overall, the Whisper-based classification model already achieved acceptable results, but it still had difficulties with certain speakers and dialect regions. Furthermore, the model tended to wrongly classify samples from speakers who came from the border areas of the dialect regions where the dialect features probably were less distinguishable.

In a second attempt we pursued an approach that has been explored very little in the context of the dialect identification task. To our knowledge, it has only been used once before for this task. The main idea behind it is to eliminate all non-linguistic features such as speaker features or noise that could distract a model from learning the dialect features, by first automatically transcribing the speech samples to phoneme sequences using a phoneme recognizer model. After that a classifier model is trained on the phoneme sequences to identify the dialect regions. To find the best performing combination of phoneme recognizer and classifier, several models and algorithms were tried out. For the phoneme recognizer, state-of-the-art pretrained cross-lingual speech-to-phoneme models were used to generate high-quality transcriptions. For the classifier, simpler classical algorithms were compared with more complex deep learning approaches. On the one hand, this led to a much more efficient training process in comparison to the first approach, since the speech samples had to be transcribed only once per phoneme recognizer, and the classifier only had to process phoneme sequences instead of raw audio data. On the other hand, the best model combination outperformed the first attempt.