Whisper Fine-tuning for Swiss German: A Data Perspective

Claudio Paonessa and **Vincenzo Timmel** and **Manfred Vogel** and **Daniel Perruchoud**
claudio.paonessa@fhnw.ch

Abstract

In our recent exploration of fine-tuning OpenAI's Whisper speech-to-text model for Swiss German, we built a data processing pipeline to transform readily available sentence-level datasets to long-form audio to be fully compatible with the Whisper model. Our pipeline ensures the preservation of the segmentation capabilities of the model and prevents the model from losing its ability to handle audio with arbitrary length. With additional High German data to preserve the German language and weakly-labeled real long-form data, annotated through the original Whisper Large V2 model, we achieve a new state-of-the-art (SOTA) model for Swiss German speech to High German text translation. The original OpenAI Whisper model shows large variations in performance across the different Swiss dialects, ranging from WER of 17.63 for Central Switzerland to 29.31 for the Valais dialect. Our model significantly improves those error rates and we measure a much more narrow range from 10.73 for Central Switzerland to 13.68 for the Bern dialect. To evaluate its performance on real long-form audio, we curated a test dataset from Swiss German TV shows with human-annotated subtitles. The novel test dataset consists of 20 hours of material from selected TV shows, i.e., Einstein, Puls, Impact Investigativ, SRF Kids News, and SRF ohne Limit.