Near-Verbatim Transcription of Dutch Speech and Its Effect on Automatic Speech Recognition

Bastiaan Tamm

Laboratory for Cognitive Neurology (LCN), Department of Neurosciences, KU Leuven, Belgium

Mara Barberis

Experimental otorhinolaryngology (ExpORL), Department of Neurosciences, KU Leuven, Leuven, Belgium

Rik Vandenberghe

Laboratory for Cognitive Neurology (LCN), Department of Neurosciences, KU Leuven, Belgium

Hugo Van hamme

Processing Speech and Images (PSI), Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

In this work, a protocol for near-verbatim transcription of natural Dutch speech is proposed. This protocol is based on the Protocol for Orthographic Transcription, which was created for the Dutch Spoken Corpus (abbreviated as CGN from the Dutch name Corpus Gesproken Nederlands), but has been adapted with the aim of improving downstream natural language processing (NLP) tasks. Applying NLP techniques to natural speech poses many challenges: for example, processing utterances with unfinished words or sentences, incorrect grammar, or non-standard words. This work focuses on the role of non-standard words and how the normalization of these words during automatic speech recognition (ASR) can help downstream applications. First, we provide the annotation protocol and normalization protocol. Next, we train end-to-end ASR models on verbatim (CGN), near-verbatim (in-house datasets) and subtitle (from Flemish TV) data using the ESPnet speech processing toolkit. We analyze the performance both quantitatively, using the word error rate (WER) metric, and qualitatively with a selection of generated transcripts. We also compare our models to the multilingually trained state-of-the-art Whisper model. Next, we perform a performance saturation analysis to estimate the amount of near-verbatim data required to finetune an existing end-to-end ASR model. This involves finetuning the model on increasingly larger splits of the near-verbatim dataset and plotting the model performance with respect to the split size. Finally, we assess the downstream performance of each model's transcripts and the reference transcript on part-of-speech tagging using the CGN dataset.