Abstract

The goal of this thesis is to develop a complete pipeline of Automatic Speech recognition for the Czech language with a particular focus on effective adaptation of the model across a variety of diverse domains. Due to the scarcity of training data, we introduce two approaches for data preparation. First, we segment a portion of our audio files in a fully unsupervised way and use them to train our baseline acoustic model. We then use this model for further refinement of the segments. With our data pipeline, we prepare over 1500 hours of training data for the Czech language, from which 444 hours are made available to the public under a non-restrictive license.

For our experiments, we use the hybrid acoustic model that combines the Gaussian Mixture Model and Hidden Markov Model with Neural Network-based methods. We also present our approach to language modeling in which we hierarchically combine interpolated n-gram models and a recurrent neural network model used to re-score the output lattices. Experiments with acoustic adaptation, which finetune the neural network to a small amount of target domain audios, are presented as well. Lastly, we introduce an efficient implementation of a model for sentence embeddings, which we use to query an extensive corpus database and condition the search on a small subset of sentences that are semantically similar to our target domain. We then use the extracted texts for language model adaptation and augmentation of the acoustic model lexicon. We demonstrate the described methods on five distinct test set benchmarks where the baseline model is compared against its adapted version as well as against the Google Cloud Czech ASR and speech recognition model developed at the University of West Bohemia. We illustrate the usefulness of the adaptation techniques by a detailed analysis of the obtained results and examination of the word error rate and domain word recognition metrics. We also present our final model, which outperforms all the proposed baselines on all but one of the considered test sets.