Iterative Improving of Transcribed Speech Recordings Exploiting Listeners' Feedback

Abstract

This Ph.D. thesis deals with making a corpus of audio recordings of a single speaker accessible to wide public and interested community.

The work has been motivated by the existence of a set of perishing recordings of the Czech philosopher Karel Makoň on magnetophone tapes. The aim is to conserve the material for future generations and to make it accessible using digital technologies, in particular publishing the recordings online and enabling the users to search through them.

The thesis introduces the creation of a system for transcribing a large set of speech recordings employing a lay community. The solution designed is based on obtaining a baseline low-quality transcription by means of automated speech recognition and developing an application that allows for collecting corrections of the automatic transcription in a fashion that makes it usable as training data for further improvement of said transcription.

The spoken corpus itself is described. The author and his works, topics covered in the talks, the process of recording and digitization as well as the gained transcription are introduced. Next, the development of a system for automated transcription of the corpus, from collecting data, to acoustic and language modeling, various experiments undertaken and evaluation are presented. Then, the web application for gathering manual transcript corrections is described. Differences to other settings, design and implementation details, a way to compensate high demand for transcription quality and low demand for worker expertise, as well as an evaluation of the system's performance after nine years of operation are covered.