Syllabify and Conquer: A report on preparing and analysing speech-text corpora

Abstract

This report documents work undertaken over a semester in the MCQLL Lab to experiment with existing speech data software tools: PolyglotDB and Montreal Forced Aligner. The experiments focused on the data preparation problems of corpora storage and speech-text alignment. The steps taken to use these tools for facilitation of a pipeline, from publicly-available corpora acquisition to concrete linguistic analysis, is described technically, demonstrated experimentally, and documented in this report.

Case studies cover the preparation of a two different tonal languages for alignment, as well as experimental studies using an already aligned corpora. This report constitutes a technical description of work undertaken using these corpora with Montreal Forced Aligner and PolyglotDB for reproduction, reuse, and as a basis for streamlining the investigation of more advanced linguistics problems that require similar speech data preparation and analysis techniques.

This report additionally describes the software design approaches taken throughout the experimental process to motivate the contribution of the set of tools written by the author and to document their workings for future users, experimenters and developers. These tools facilitate the generalization of the experimental procedure using languages and corpora beyond those covered in the report experiments. This generalization is accomplished by abstracting the differences between corpora to a set of configuration files adaptable to the requirements of large sets of known public corpora. Differences in corpora that cannot be abstracted to configuration are implemented using a Strategy Pattern-based approach, streamlining the introduction of new client logic for new corpora to the adaptation of a testable template class. This approach aims to contribute a code-base that can dynamically adapt to and reliably implement future client requirements when new corpora necessitate novel business logic in configuration and linguistic structure.

Overall, this report describes an approach and some software solutions to aid in the resolution, by non-expert and expert researchers alike, of preparing publicly available speech data for alignment, storage, and analysis using publicly available tools.

Submitted: June 15, 2022
Author: Michael Haaf ***(michael.haaf@mail.mcgill.ca)***
Supervisor: Morgan Sonderegger ***(morgan.sonderegger@mcgill.ca)***