Extending the EMU Speech Database Management System: Cloud Hosting, Team Collaboration, Automatic Revision Control

Markus Jochim[1]

[1]Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität München, Germany

markusjochim@phonetik.uni-muenchen.de

Abstract

In this paper, we introduce a new component of the EMU Speech Database Management System [1, 2] to improve the team workflow of handling production data (both acoustic and physiological) in phonetics and the speech sciences. It is named *emuDB Manager*, and it facilitates the coordination of team efforts, possibly distributed over several nations, by introducing automatic revision control (based on Git), cloud hosting (in private clouds provided by the researchers themselves or a third party), by keeping track of which parts of the database have already been edited (and by whom), and by centrally collecting and making searchable the notes made during the edit process.

Index Terms: speech databases, cloud hosting, groupware, team coordination, phonetics, data management, workflow

1. Introduction

The software tool presented in this paper, the *emuDB Manager*, facilitates important steps in the research workflow of handling speech production data. Managing speech production data involves the collection, segmentation, annotation, and analysis of speech. These steps require many hours of manual labor and result in copious amounts of primary as well as secondary data files. To aid researchers – and especially teams of researchers – with their task, emuDB Manager allows them to systematically share their files on a server, keep track of different versions, keep track of their notes, and coordinate which parts of the data have already been edited (and by whom) and which parts remain to be edited (and by whom).

2. EMU Speech Database Management System

The EMU Speech Database Management System (in short EMU-SDMS) mainly comprises a web application to visualize, annotate and segment data, and two R packages to analyze data, with a sophisticated query language to allow for the analysis of database subsets based on their hierarchical and sequential annotation. The latest developments have been described in [1]. The system's website is found at [3].

The emuDB Manager adds to this a server component for centrally storing data and assigning tasks to project members; and a web frontend to control these functions and to upload/download data as necessary. The server component extends an already existing reference implementation of the EMU-webApp protocol.

Researchers wishing to host an "EMU cloud" at their own institution need to run a web server and a nodeJS server. The source code of all components is found on GitHub [4].

3. Cloud Hosting

Cloud hosting basically means to store data (primary and secondary) on a central server, such that all team members can access and modify them (after being authenticated through their password). Cloud hosting can greatly facilitate collaboration in teams, no matter if they are located at a single university or spread over research institutions world-wide. The key advantage is that team members do not need to store portions of the data on their own computers, edit (segment and annotate) them in parallel and then put the individual portions back together. Editing is done directly on the server-stored data. This avoids much confusion about which copy is the most recent one or the "master," since the server ("the cloud") holds all modifications.

The emuDB Manager web frontend is used to upload data onto the server, where all team members can see them. The EMU-webApp can then be used to visualize and to edit the data. Once the edit process is finished or has progressed sufficiently to start analysis, the edited data can be downloaded again. The version being downloaded can also be given a label, which is especially handy for intermediary analyses (in the middle of the edit process). By referring to the labels, the team can always be sure which version of the data has been analyzed.

Data protection is a very important aspect of cloud hosting. It is in the interest of both researchers and participants that the data not become public before the researchers decide to publish them in a way that upholds the participants' anonymity and other legal rights. The emuDB Manager achieves this by requiring users to authenticate with their password before they gain access to any data. The project leader can choose accounts that are entitled to access and modify the project data. For this reason, the project leader can create new accounts or utilize ones provided by research institutions (using Shibboleth), e. g. via the CLARIN Service Provider Federation [5]. This is useful for teams comprising multiple universities, since each member can use the account they already possess.

Data protection also extends to the question where the data is physically stored, and that can be on a server located at an affiliated research institution (as is the case for the author's institute) or on a commercially rented server.

4. Automatic Revision Control

Over the course of a project, data is added and edited by several persons. The first analysis steps are often started before the edit process has been completed, and sometimes analysis reveals that more collection is due or that annotation guidelines need changing. This makes it necessary to consistently distinguish different versions of the whole data set and assign labels to these versions. The software Git provides a very stable mechanism for this. The emuDB Manager uses Git to store a snapshot of the data whenever data have been added or edited, and when work has been (re-)assigned to collaborators. This way,

no changes are left undocumented. It is transparent at any time which changes have been made, when, and by whom. Any set of changes (called a *commit*) down to the resegmentation of an individual recording can be traced to its editor and time, and if need be, it can be reversed.

5. Team Collaboration

Coordination of these tasks becomes much more difficult as the number of collaborators and assistants grows, and also as the amount of data grows. emuDB Manager provides two tools for efficient coordination. Research teams can decide whether to grant full access to everybody including assistants, or restrict the assistants' access to those parts specifically assigned to them.

5.1. Whose work is it?

The first tool is that team members can assign their colleagues (or self-assign) parts of the database for editing. This is done using *bundle lists*.[1] To create a bundle list, researchers select a part of the database using regular expressions (or the whole database) and assign them to one or more team members. When team members log in for editing, they will only see the parts currently assigned to them.

All bundle lists are gathered in a clear overview, indicating the percentage to which they have been finished. This enables users to see at a glance who is currently editing which files and how far work has progressed.

Once a bundle list has been finished, it is given an archive label. It will still be shown in the overview, but clearly marked as "finished work." The archive label can be any text freely chosen by the team members. While the software imposes no restrictions, useful choices may include the date when the bundle list was finished, or a name for the portion of work or editing stage that the bundle list represents.

5.2. Collecting Notes

Many times, especially while segmenting and annotating, researchers take notes about particular recordings. emuDB Manager allows researchers to store these notes centrally, and makes it possible for every team member to search these notes.

Notes can be entered in EMU-webApp, the same interface that is used for editing and visualizing data. Notes will be stored along with the bundle lists. In emuDB Manager, researchers can read and search these comments. It is also possible to restrict the view to "commented bundles only," and then use EMU-webApp for visualizing the respective bundle data (spectrogram, segmentation, formants etc.), along with the comments.

This is very helpful especially for the communication between project collaborators and student assistants, when discussing the annotation of individual recordings.

6. Discussion

emuDB Manager is an extension of the EMU Speech Database Management System (EMU-SDMS). EMU-SDMS "sets out to be as close to an all-in-one solution for generating, manipulating, querying, analyzing and managing speech databases as possible." [1] This new extension adds several features that are especially valuable when working in teams: It provides speech scientists with an easy-to-use interface to established (but often complicated-to-use) techniques such as automatic revision control or server storage; further, it also facilitates usage of advanced features of the EMU-SDMS, such as bundle lists.

The most interesting next steps are concerned with further exploiting cloud hosting and cloud computing, as well as with inter-labeler agreement. As to inter-labeler agreement, the emuDB Manager already forms a useful tool for multiple editors to control each other's work. A useful extension to this would be to automatically evaluate inter-labeler agreement when multiple editors work on the same recordings independently of each other.

As to cloud functionality, we are considering two aspects. Since data analysis is carried out with the R package emuR, one of the two is to exploit RStudio's server version to bring the analysis step into the cloud. Currently, it is necessary to download the data from emuDB Manager in order to perform analyses, once the edit process is finished. RStudio in the browser would therefore yield two advantages: Server-side processing power could be utilized for computation-heavy statistical analyses (cloud computing); and the need for the error-prone task of repeatedly copying large data sets between machines would be further reduced.

The second aspect is who provides the cloud resources. While each research institution could use the software (which is freely available under an open source license [4]) and run a cloud service for its own members, it might be more efficient if laboratories shared their resources. The CLARIN Center BAS [6], hosted by the University of Munich, already provides speech processing tools to the research community (e. g. WebMAUS for automatic segmentation) and is considering to include the emuDB Manager in its cloud services. It may also be worthwhile to integrate EMU-SDMS with infrastructures such as the Open Science Framework [7], or with commercial cloud service providers.

The emuDB Manager is already proving a valuable tool for our international team spread over three research institutions and we hope that in the future, other groups will be able to share the same advantages.

[1] In EMU-SDMS, a recording and its accompanying secondary files (e. g. annotation, derived signals) are termed a "bundle."

7. Acknowledgements

This work was supported by the DFG-DACH grant number KL 2697/1-1 "Typology of Vowel and Consonant Quantity in Southern German varieties: acoustic, perception, and articulatory analyses of adult and child speakers" awarded to F. Kleber.

8. References

[1] R. Winkelmann, J. Harrington, and K. Jänsch, "EMU-SDMS: Advanced speech database management and analysis in R," *Computer Speech & Language*, in press.

[2] R. Winkelmann, "Managing speech databases with emuR and the EMU-webApp," in *INTERSPEECH-2015*, 2015, pp. 2611–2612.

[3] Institute of Phonetics and Speech Processing. The EMU Speech Database Management System (EMU-SDMS). [Online]. Available: https://ips-lmu.github.io/EMU.html

[4] ——. GitHub Projects. [Online]. Available: https://github.com/ips-lmu

[5] CLARIN ERIC. Service provider federation. [Online]. Available: https://www.clarin.eu/content/service-provider-federation

[6] Bavarian Archive for Speech Signals. [Online]. Available: http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B

[7] Open Science Framework. [Online]. Available: https://www.osf.io/