WikiSpeech – A Content Management System for Speech Databases

Christoph Draxler, Klaus Jänsch

BAS Bavarian Archive for Speech Signals
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München, Munich, Germany
{draxler|klausj}@phonetik.uni-muenchen.de

Abstract

In this paper we describe WikiSpeech, a content management system for the web-based creation of speech databases for the development of spoken language technology and basic research. Its main features are full support for the typical recording, annotation and project administration workflow, easy editing of the speech content, plus a fully localizable user interface.

For the creation of a new speech database, it is only necessary to open a new project within WikiSpeech, provide a link to any static project information pages and upload the prompt material to be presented to the speakers. Recordings and annotation are performed via the WWW in a platform independent manner on any Java compatible computer.

WikiSpeech currently has been localized to four languages: German, English, Romanian and Russian, and it is now used for production recordings at the Bavarian Archive for Speech Signals in Munich, Germany.

Index Terms: speech database collection, web-based recording, web-based annotation, content management systems, internationalization

1. Introduction

In the last few years, web-based systems for recording or annotating speech have been developed to support the basic tasks required for the creation of speech resources and education, e.g. [1, 2]. These systems include SpeechRecorder for recording speech [3], and WWWTranscribe [4], Transcriber [5], Web Transcription Tool [6], Phonex [7], WebTranscribe [8], and others for the annotation of the recorded speech signals.

More recently, web-applications have been built that combine both recording and annotation components to provide an integrated environment for the development of large speech databases. For example, a web application was developed at BAS for the German Ph@ttSessionz speech database which contains read and spontaneous speech of adolescent speakers collected in public schools more than 40 recording locations in Germany [9]. In a similar vein, the SPICE system allows the rapid web-based creation of speech resources for the development of speech recognizers [10]. These web-applications perform well on the task for which they were designed, but they are quite difficult to adapt to new speech database collections, and a new approach is needed to provide more flexibility.

Two web-technologies offer this additional flexibility: content management systems and wiki systems.

A content management system (CMS) is a web application that implements a given data model, provides a graphical user interface and supports the workflow of a given task. Examples of CMS are

- courseware systems like Moodle for e-learning with a data model for course creators, teachers, students, lectures, exams and peer-to-peer and student-teacher communication,
- e-shopping systems with a data model for merchandise, catalog, shopping-cart, and payment, or
- forums and blog sites with a data model for authors, topics, trails, message exchange, etc.

A wiki system is an interactive web site, where all users can also be authors of the sites pages, so that an update of a page is immediately visible to all other users. The prime example of a wiki site is, of course, Wikipedia, the online encyclopedia created an maintained by its user community.

Section 1 presents the system design of WikiSpeech, section 2 the supported configurations. Section 3 gives a walkthrough example for setting up a speech database collection for a new resource creation project.

2. Architecture of WikiSpeech

WikiSpeech is a CMS for the creation of speech resources. It implements a data model and the application logic for speech database creation, provides a localizable graphical user interface, and allows a user-defined specification of new data collection projects.

The creation of speech resources consists of four main tasks: Database specification, speech recording, annotation and project administration. In WikiSpeech, these tasks are performed via the web on distributed clients connected to the server. A WikiSpeech site can perform many speech resource creation projects in parallel. Within each such project, the individual tasks can also be run in parallel.

For specifying a speech database collection project, a database author enters the project language and recording parameters via a web form, and uploads the recording scripts to the server.

For recording, a speaker selects a recording project from a list presented on the screen. He or she then enters his or her demographic data via a web form. This data consists at least of the speaker age, sex, native language, city or region where he or she entered school. Weight and height and other potentially speech relevant personal traits are optional, e.g. smoking habits, dental braces, or piercings in the tongue or lips (see fig. 3 for a sample form). Once this data has been entered, the server starts a recording session and presents records one prompt item after the other until the session is done.

For annotation, the administrator assigns annotators to a given project. The annotators log in, select a recording session to annotate, and process one signal file after the other. Once

an annotation is done, it is uploaded to the server and the next recording to be annotated is selected.

Project administration provides an overview of status of a data collection project: querying the database, monitoring of ongoing recording sessions, quick access to signal files for quality control, communication with speakers or database authors, statistics on the annotation, etc.

2.1. Web Application

A web application is an enhanced client-server system in which the server provides the application logic and the data storage, and the client implements the user interface and some data processing, e.g. form validation or signal processing.

Data exchange between server and client is perfomed either using data objects implemented natively in the programmg language of the web application, or in an exchange format, usually XML.

The main difference to the common client-server architecture is that the server of a web application implements session management. Session managment keeps track of the state of the individual clients, e.g. initializing a recording session, status of the data upload, terminating a session, etc., and monitors the general project progress.

2.2. Data model

The data model of WikiSpeech is the outcome of the experiences gained in the German Ph@ttSessionz data collection. A rudimentary version of this data model was the basis for the Ph@ttSessionz database. In the course of the project, the data model was extended incrementally to meet the growing demands. For WikiSpeech, the data model was refined and many project specific details were modelled in a more abstract way to allow an easy adaptation to other speech database collections.

2.2.1. User types

WikiSpeech distinguishes four classes of users: *Speaker*, *Technician*, *Annotator* and *Administrator*, all of which are subclasses of *Person*. They thus share all the attributes of *Person* and extend this class by additional attributes, e.g. demographic data for a speaker, technical qualification of a technician, the formal training of an annotator, and access privileges for the administrator.

The administrator creates a project, adds and edits the recording scripts, grants or revokes access rights to technicians and annotators, and assigns annotators to annotation projects. Finally, the adminstrator supervises the progress of each recording and annotation session.

2.2.2. Classes and relationships

The data model of WikiSpeech is given in fig. 1. It distinguishes the following classes and their relationships: an *Organization* performs recording *Project*s and *AnnotationProject*s. A recording *Project* consists of *Session*s. A session is an organizational unit comprising a *RecordingScript* and recording equipment, and it is supervised by a technician. The recording script consists of *Section*s which in turn contain *Recording*s. A speaker performs a *Recording* and thus produces one or more *Signal*s, stored in an audio file.

An organization, not necessarily identical to the recording organization, specifies an *AnnotationProject*. This consists of *AnnotationSession*s which associate annotators and *Annotation*s. An *Annotation* describes the content of a *Recording* on a given annotation tier and in a given annotation format.

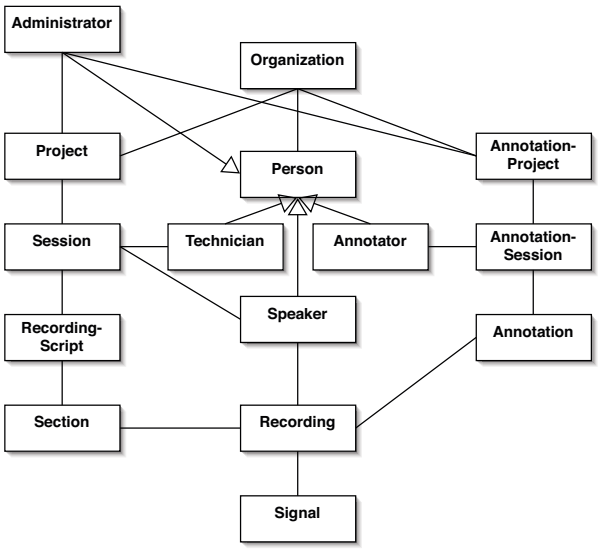

Figure 1: *Simplified class diagram of the WikiSpeech data model*

2.3. Components

WikiSpeech integrates the applications SpeechRecorder for recording speech [3] and WebTranscribe [8] for the annotation of speech as components.

SpeechRecorder is a Java Web Start application that allows recording speech via the Internet. Prompt items are downloaded from the server and displayed on the screen to be read or answered by the speaker. SpeechRecorder not only supports Unicode text prompts, but also image and audio prompts to elicit spontaneous speech or perform 'speaker after me' recordings. The recording capabilities depend on the hardware connected. The number of channels, sample rate, quantization and compression are set via the project settings on the server. All signal files are transferred to the server during a recording session in a background process; if the client is shut down before all data has been transferred, SpeechRecorder makes sure the remaining data is transferred to the server the next time it is run again.

For annotation, the annotation framework WebTranscribe is used. WebTranscribe implements the 'select-annotate-save' workflow for speech annotation and features a signal display and an annotation editor. This editor provides the functionality needed for the given type of annotation, and it is implemented as a plug-in module. Plug-ins have been developed for e.g. a basic orthographic annotation with noise and signal quality markers as used in the SpeechDat project [11], for annotating dysarthric patients by selecting words from a list [12], and others.

2.4. System requirements

WikiSpeech is intended to be used at speech resource creation centers. Such centers provide the web space necessary to store large amounts of speech data, and have system administrators for the installation and setup of Java web applications and database systems.

WikiSpeech is implemented in Java version 5. The web pages are dynamically created using Java Server Pages (JSP), a mature and widely used technology.

Figure 2: *Localized main menus of WikiSpeech*

On the server side, WikiSpeech requires a relational database management system (RDBMS), e.g. PostgreSQL, mySQL, or Oracle, and a web server capable of serving JSP, e.g. Tomcat. The RDBMS is used to store the user data, speech content, session data and other symbolic data. The speech signals are normally stored in the server's file system. For enhanced security, the speech signals can be stored in the database system.

The client must have a virtual machine with Java version 5 or newer. The recording and annotation applications are implemented using Java web start. This means that all applications are automatically downloaded from the web, installed and run with privileges of the current user. During recording, signal data is stored on the client in temporary buffer files – once the data is uploaded to the server, no data remains on the client machine.

3. Creating a speech database project

3.1. Language adaptation

WikiSpeech distinguishes two levels of language adaptation: user interface localization for the web interface, and speech database localization for the database contents.

3.1.1. User interface localization

The user interface of WikiSpeech consists of JSP pages with placeholders for localizable content. Upon loading the pages, the web server replaces the placeholders with user interface texts corresponding to the project language. Fig. 2 shows the main menu for German, English, Russian and Romanian, and fig. 3 shows localized versions of the forms for entering demographic speaker data and project administration.

Adapting the graphical user interface of the WikiSpeech web application to a new language requires new translations for the placeholders in the WikiSpeech resource files. These translations are entered via a web form. This form contains a complete and uneditable list of placeholders, and for each placeholder an input field for the translated text.

Once all placeholders have been processed and checked for completeness and consistency, the system administrator halts the WikiSpeech server and adds the newly generated resource files to the application software. This bundle is then redeployed

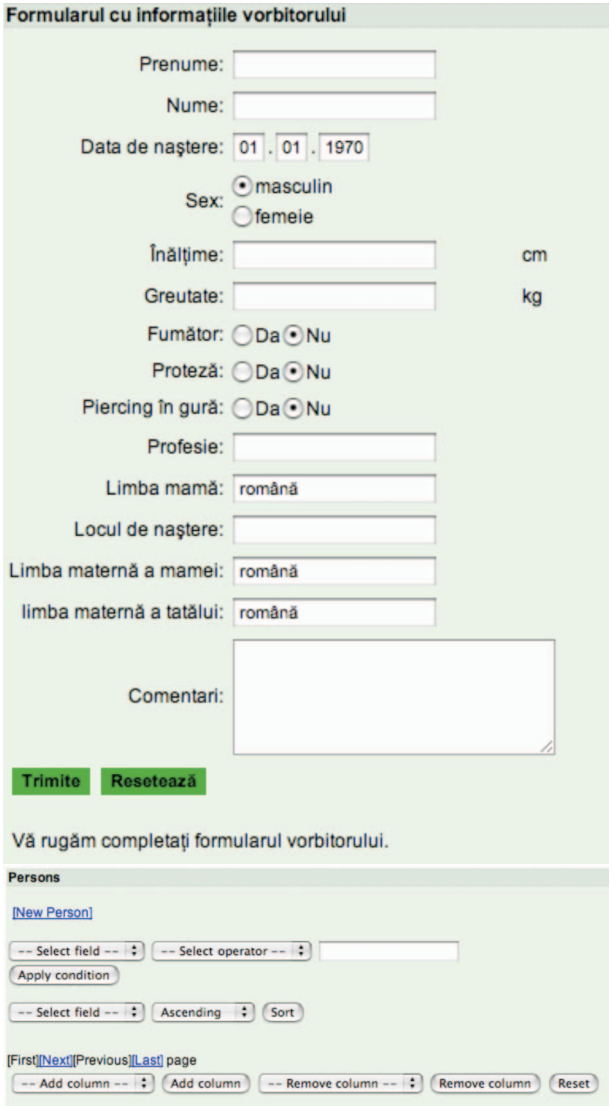

Figure 3: *Romanian form for entering demographic data, and English query form for searching the person database*

on the server. During this period the system is not available to users, but this procedure normally takes only a few seconds, and it is a rare event.

3.1.2. Speech database contents specification

The speech database contents are defined in recording scripts which are organized in sessions. Each session uses a single recording script, and a given recording script may be used in more than one session.

Recording scripts are XML files. Fig. 4 shows a short recording script with a single section and two recording items. A database author creates recording scripts outside WikiSpeech using an XML editor and uploads the scripts to the server. The server checks the script for completeness and consistency before storing it in the relational database system.

```
<recordingscript>
  <section name="Intro" order="sequential" speakerdisplay="off"
           mode="manual" promptphase="idle">
    <recording prerecdelay="1000" recduration="5000" postrecdelay="500" itemcode="Z0">
      <recinstructions mimetype="text/UTF-8">
        Here are the instructions...
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          ... and here's the text to read.
        <mediaitem />
      </recprompt>
    </recording>

    <recording prerecdelay="1000" recduration="5000" postrecdelay="500" itemcode="Z1">
      <recinstructions mimetype="text/UTF-8">
        Please read
      </recinstructions>
      <recprompt>
        <mediaitem mimetype="text/UTF-8">
          Adjust the level until the signal is clearly visible.
        </mediaitem>
      </recprompt>
    </recording>
  </section>
</recordingscript>
```

Figure 4: *Sample recording script with English prompts.*

3.2. Recording parameters

Finally, the project administrator specifies the project signal quality settings via a web form. These settings include sample rate, quantization, encoding and file format – currently, only WAV is supported for audio recordings. Note that WikiSpeech passes these to the SpeechRecorder component. If the audio hardware connected for the recordings does not match these settings, then recordings will fail with an error message.

4. Conclusion and Outlook

The first version of WikiSpeech was installed at BAS in March 2008 with localization available for German, English, Romanian, and Russian. The current version (June 2008) is quite stable already, but presently not all services are available – selecting a recording project, input of demographic speaker data ad performing the recordings are fully functional. For annotation, only a basic orthographic transcription editor and a word-list selection editor are implemented.

The first data collection project to run on WikiSpeech is VOYS, a collaboration between BAS and QMU Edinburgh, scheduled to start in late 2008. In VOYS, speech will be recorded in public schools in 10 locations in Scotland, and the data will be stored on the server at BAS.

BAS offers to host additional speech database projects, and it provides support and service for setting up and running these projects. For details, visit `www.wikispeech.org`

5. Acknowledgments

We gratefully acknowledge the contribution of Marlis Friedl, Monyka Tarnu and Fatima Usmanova of the Sprachen- and Dolmetscher Institut München who provided the localization of the WikiSpeech user interface and database contents for Romanian and Russian.

6. References

[1] Sjölander K., Beskow J., Gustafson J., Lewin E., Carlson R., Granström B., "Web-based Educational Tools for Speech Technology", Proc. of ICSLP, Sydney, 1998, pages 3217-3220

[2] Drygaljo A., Delafontaine G., "Using Java to Develop Interactive Learning Work-Benches for Speech Analysis Basics on the World-Wide Web", Proc. of MATISSE, London, 1999

[3] Draxler Chr., Jänsch K., "SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software", LREC 2004, Lisbon.

[4] Draxler Chr., "WWWTranscribe – A Modular Transcription System based on the World Wide Web", Proc. of Eurospeech, Rhodes, 1997

[5] Maidment J., "Transcriber", `http://www.btinternet-.com/ eptotd/vm/transcriber`, last checked 09/Apr/2008

[6] Garcia Lecumberri M., Maidment J., Cooke M., Ericsson A., Mircea G., "A web-based transcription tool", Proc. of ICPhS, Barcelona, 2003

[7] Jensen Chr., "Online Training and Testing in Phonetics", Proc. of PLTC, London, 2005

[8] Draxler Chr., "WebTranscribe – An Extensible Web-Based Speech Annotation Framework", Proc. of TSD, Karlsbad, 2005

[9] Draxler Chr., Jänsch K., Speech Recordings in Public Schools in Germany - the Perfect Show Case for Web-based Recordings and Annotation, Proc. of LREC 2006, Genova.

[10] Schultz T., Black A., Badaskar S., Hornyak M., Kominek J., "SPICE: Web-based Tools for Rapid Language Adapation in Speech Processing Systems", Proc. of Interspeech, Antwerp, 2007

[11] Winski R., "Definition of Corpus, Scripts, and Standards for Fixed Networks", SpeechDat Report LE2-4001-SD1.1.1, 1997.

[12] Ziegler W., Hartmann E., "Das Münchner Verständlichkeitsprofil (MVP) – Untersuchungen zur Reliabilität und Validität, Nervenarzt 64, pp. 653-658, 1993