A Web based Speech Transcription Workplace

Markus Klehr, Andreas Ratzka, Thomas Roß

European Media Laboratory GmbH, Heidelberg, Germany
{markus.klehr|andreas.ratzka|thomas.ross}@eml-d.villa-bosch.de

Abstract
We describe our web based speech transcription tool EML Transcription Workplace (TWP). Apart from its main purpose of annotating audio data, it also includes support for the management of transcription data, ASR based pre-transcription, assignment of work packages to specific users, user management and a correction/verification workflow. These features help to increase the productivity for both transcriptionists and supervisors and facilitate further processing.

Index Terms: speech, transcription, data management, workflow

1. Introduction
To improve the quality of automatic speech recognition, language and acoustic models need to be trained on representative data. In order to cover also spontaneous speech (from interviews, talk shows, call center dialogs, voice mails) real life application data is indispensable as spontaneous speech differs from written language both syntactically and in terms of the words used [6]. Therefore, to get realistic data for training, it is necessary to manually transcribe audio data.

Standalone tools (e.g. Transcriber [2]) have two intrinsic drawbacks if transcriptionists work on more than a few machines or work remotely:

- Installation: The transcription software usually supports only a fixed number of platforms and has to be installed on each machine being used for transcription separately.
- Data Management: Audio data needs to be distributed to transcriptionists and the transcription output produced by multiple transcriptionists has to be collected and merged into central repositories.

To reduce this overhead we developed the web-based EML Transcription Workplace (TWP). The graphical user interface runs in a web browser. This means, no further installation is required on the user side. The data (audio data, transcriptions, speaker information, dictionaries for spell checking, ...) are stored in a central repository on a server. The application supports data management and job assignment and users are not required to maintain local copies of the audio and transcription data. Furthermore, a review and correction cycle is supported to simplify quality checks and the re-assignment of transcription jobs.

2. Architecture
TWP is a web application and runs inside a JEE [7] web container of the geronimo application server 2.2.1 [1]. For the graphical user interface the JavaScript based YUI 2.8 library is used [11]. To simplify the client-server communication with JavaScript, we use direct web remote (dwr) [3]. The application is tested for the browsers Firefox and Chrome. The audio data playback uses the JW Flash player (client side) [9] and the Red5 streaming server (server side) [8]. With this setup, audio streaming is used and no data needs to be downloaded and stored locally.

![Figure 1: EML Transcription Workplace Architecture](image)

For an intuitive representation of the audio data, the user interface requires some auxiliary files that are created automatically during the data import: Streamable audio files that are compatible to the Flash Player and the Red5 [8] streaming server (e.g. speex [10] encoded audio in a flv container) and waveform visualisations to enable navigation in the audio signal.

If a suitable speech recognition model is available it might be helpful to produce automatic pre-transcriptions...
to speed up the transcription process. For this, sending audio data automatically to the EML Transcription Server [4] is supported.

3. Transcription Tool Features

3.1. Speech Segments

Some speech recognition tools restrict the length of input audio to a maximum length. This means, long audio files need to be organized into shorter segments consisting of start and end times plus the transcription of the respective segment. Also, several steps in acoustic model training often require speaker information to be available on a segment level (VTILN, SAT). Other segment level information includes background noise, music, unintelligible speech etc. All the segment information listed above can be stored using TWP. Figure 2 shows currently supported annotations and how they are displayed to the user.

![Figure 2: Markup example on segment level](image)

If dialogs with overlapping speech are recorded with multiple channels, the user can select to playback a single track (for precise transcription) or all tracks at once (to get an overview).

3.2. Spell Checking

When transcriptions are manually created or corrected, text data is uploaded to the server and then processed by a spell checker (Hunspell [5]) to eliminate spelling errors. With TWP, the transcriptionist can only commit transcriptions containing words which are not covered by the spell check dictionary when each unknown word is added to a personal spell check dictionary. Before new words are added to project-wide or general dictionaries, the supervisor can review these words, including personal spell check dictionaries of transcriptionists. Each user created dictionary entry is linked to its first occurrence in the transcription and can therefore easily be reviewed by the supervisor.

3.3. Pronunciation support

Colloquial variants in pronunciations of words can be provided by the transcriptionists as a soundslike spelling: The orthographically correct spelling is transcribed and then annotated with a word that, if read by a native speaker in the respective language, sounds like this variant. Although this is not a phonetically correct representation, it can be created by uneducated users and supports letter-to-sound rules to automatically produce better pronunciations for the respective word. The same process of annotation can be used to provide pronunciations for acronyms.

3.4. Automatic pre-transcription

The transcription effort can be minimised if a speech recognition model for the respective domain is available. TWP can make use of all models available in the EML Transcription Server [4]. The model selection is integrated into TWP and a suitable model for pre-transcription can be selected when the data is imported.

3.5. Access and user right management

To enable access to customers and outside partners for trial or productive use of our services and tools, a fine grained access control is required to constrain the access to specific transcription projects. LDAP based user authentication is used to implement this. Access rights are based on group membership.

3.6. Anonymisation of Audio Data

In some scenarios (e.g. data from call center or voicemail providers), data may be used only, if personal information is removed from the audio. For this, the platform supports muting confidential segments of the audio signal and replacing corresponding confidential text with class information (e.g. firstname, lastname, address, ...). The usage of class names retains information which can be used for language model training. Anonymisation is usually done on customer’s premises by transcriptionists with appropriate access rights. Anonymised data can then be further processed by external transcriptionists.

4. Usage of the Transcription Tool

If no speech recognition model is available (or the quality of available models is too low, e.g. because of a model / domain mismatch) the audio data needs to be transcribed completely manually.

4.1. Creating transcriptions from scratch

The most convenient way is to enter the editing mode, play parts of the transcription using the tabulator key, stopping the playback with the control key whenever necessary, inserting breakpoints via clicking into the audio signal. An example is shown in figure 3.

Instead of clicking into the audio signal, the users can also set marks to the last stop position via Ctrl + M and
4.2. Correcting pre-transcriptions

When the audio data is pre-transcribed the users have to correct mis-recognized words. As the recognizer returns pre-segmented data, the efforts of inserting breakpoints can be significantly reduced.

The user typically enters the edit mode for a segment, corrects misspelled words, inserts noise markers, and segment markup, pronunciation variants and acronyms. Additionally the user has to assign a speaker id to each segment. If necessary, the user can add a new speaker id to the data base. For this he has to provide gender, name, accent, role, and language information. If the language information differs from the language given at file level, the speaker language overwrites the language information used by the spell checker and by the subsequent processing steps. This way, multi-lingual files can be processed appropriately.

In case of misplaced segment boundaries (i.e. if the time alignment is incorrect), the transcriptionist can remove it using Control + Del (the following segment boundary is removed) or Control + Backspace (the previous segment boundary is removed). The transcriptionist can quickly jump to the next segment to edit and proceed his work.

When leaving the edit mode, the corrected text is marked in several ways:

- Text that was inserted is underlined with green dots.
- Incorrect words spotted by the spell checker are underlined in red.
- Words the user has added to the spellchecker are underlined in blue dots.

Before committing the transcription the user has to correct the words spotted by the spell checker. For this he can select suggestions from a menu, add the new word to the spell check dictionary or edit the text again. After committing the transcription the next audio to be transcribed is loaded automatically.

4.3. Anonymisation of confidential data

Audio data with confidential content (e.g. names, phone numbers, PIN codes etc.) may require additional processing before being passed to non-authorized people. For this, TWP supports two approaches for anonymisation.

- If a high accuracy model is available for the respective domain, it might be sufficient to automatically replace confidential words in the recognizer’s output by corresponding class information and muting the respective segments in the audio signal. The time information for muting is provided in the recognition output. Confidential words must be provided in separate lists, one list per category (person names, street names etc.).
- Otherwise, the transcriptionist will listen to the audio and select those words in the uncorrected output which correspond to the confidential parts in the audio and tags these with the respective class information. This semi-automatic process does not require time consuming navigation in the audio signal.

After tagging, the supervisor starts the anonymisation process which mutes the confidential audio and deletes the original data. The confidential parts of the text are replaced by class information.

5. Administration, Data Organization and Workflow

A major aspect in developing TWP was the support of remote access and remote work for both the transcriptionists and the supervisor. The separation of data into projects and the integration of LDAP based user authentication allows for a fine grained organization of transcription data.

5.1. Workflow

The usual cycle of data transcription consists of data import, job assignment, transcription, and review.

5.2. Import

The data in TWP is organized in projects which are defined by a) a project name, b) a customer name, c) a comment field, d) a default ASR model for pre-transcription, e) an anonymisation flag, f) the user group that may work on this project, and g) the default language of this project. Single audio files or zip-archives may be imported. It is also possible to import pre-transcriptions that conform to a specific XML format. Only the supervisor may create new projects or import new media (audio files and – if available – pre-transcribed text).

If importing audio only, automatic pre-transcription can be selected if an appropriate ASR model exists.
5.3. Job assignment

Imported data is assigned by the supervisor to individual transcriptionists or to a group of transcriptionists. For this, he can either select a complete project, an import data set or even single files in the data overview. After selection of a data set, an aggregated table is displayed which lists the duration and the file count for the following subsets: the imported data, the data assigned to each transcriptionist, the data that already has been transcribed, and the data that was already exported.

The supervisor can now select a subset in the aggregated view and delegate the corresponding data to an individual transcriptionist or to a group of transcriptionists associated with the current project. In order to control the job assignment at a finer grained level, the supervisor can limit the amount of data selected in the current step in terms of number of files or minutes of audio, respectively. The supervisor can also assign a priority to the jobs.

Only transcriptionists that are member of the corresponding user group are presented in the selection list. Users may be member in several user groups with both user and manager roles.

When a user logs in, the system looks for data assigned to this user in decreasing order of priority. Data directly assigned to the user is given precedence over data assigned to those groups the user is a member of.

5.4. Optional: Anonymisation

Data that has to be anonymised can be marked by the supervisor as confidential. If so, a word list for each confidential category can be uploaded and the recognition result is matched against the words in each category. Matching words are replaced by their respective class markers and audio segments corresponding to recognized words are muted.

5.5. Review

Completed transcriptions are committed by the transcriptionist and can then be reviewed by the supervisor. In case of errors or corrections, words or phrases can be labeled with comments and then be sent back to the transcriptionist for re-work.

The supervisor can also review personal spell check dictionaries, pronunciation variants and acronyms. All these entries are linked to the respective transcriptions and are therefore easily accessible for the supervisor for verification purposes.

6. Conclusions

We have described the basic features of our web based transcription tool EML Transcription Workplace (TWP). In addition to the core functionality of a transcription tool the support of distributed, remotely working transcriptionists has been emphasized. Workflow issues including review and correction cycles have been taken into account as well.

7. Acknowledgements

We would like to thank all our colleagues at EML who have contributed to TWP. As the pre-transcription feature makes use of all speech recognition related components, it actually means all colleagues.

8. References


Figure 4: Transcription Workflow