TSAB – Web Interface for Transcribed Speech Collections

Tanel Alumäe[1], *Ahti Kitsik*[2]

[1]Institute of Cybernetics at Tallinn University of Technology, Estonia
[2]Codehoop OU, Tallinn, Estonia
tanel.alumae@phon.ioc.ee, ahti.kitsik@codehoop.com

Abstract

This paper describes a new web interface for accessing large transcribed spoken data collections. The system uses automatic or manual time-aligned transcriptions with speaker and topic segmentation information to present structured speech data more efficiently and make accessing relevant speech data quicker. The system is independent of the underlying speech processing technology. The software is free and open-source.

Index Terms: speech collections, web interfaces, speech indexing

1. Introduction

In the recent decade, there has been an increase in the availability of online spoken audio content. Many broadcasting organizations have started to provide audio and video archives of their programs (such as discussion programs and interviews) on their web sites. Often, the recordings are provided via RSS feeds as podcasts. However, the sheer volume of such archives makes efficient access to such data difficult. Usually, the recordings are accompanied by no or little manually associated metadata. Often, the only option to discover the contents of a spoken audio document is to listen to it which takes time proportional to the length of the recording. This makes locating specific information within large spoken audio collections difficult and frustrating.

The Transcribed Speech Archive Browser (TSAB) is a web interface for automatically or manually transcribed and segmented spoken language collections. The system aims to make such digital audio archives more accessible and easier to use by employing time-aligned transcriptions of speech data to provide retrieval and browsing functions to the end-user.

The TSAB system serves as a front-end to transcribed audio archives. It is not tied to any specific archive, however. It is possible to deploy new "themed" instances of TSAB for specific purposes. For example, TSAB could be part of a conference web site and provide access to transcribed recordings of conference presentations, universities could use TSAB to provide access to their lecture recording archives. In the pilot deployment, the system serves as a front end for automatically transcribed recordings from different conversational programs from several Estonian radio stations. The system is deployed at `http://bark.phon.ioc.ee/tsab`.

The software is released as open source under the AGPL license. The source code is available at `http://code.google.com/p/tsab/`.

The software is partly inspired by a few previous systems developed for improving access to large spoken language collections. Probably the most well-known interface to automatically transcribed speech archives is the MIT Lecture Browser[1]. However, the MIT Lecture Browser is not available to be used for other kinds of speech collections, it's not open source, and it relies on the RealPlayer browser plugin that is nowadays not installed on many computers. LIMSI's AudioSurf[2] had also similar goals as our system, but it is not web-based, not open, and its development seems to have stopped.

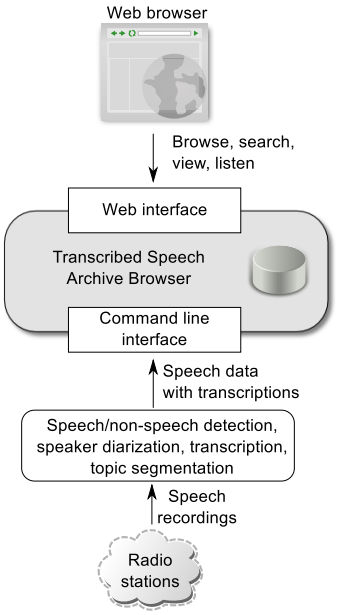

Figure 1: Interconnection diagram of the TSAB system.

2. Architecture and features

The context of the TSAB system is outlined on Figure 1. TSAB runs on a web server and maintains an index of speech recordings and corresponding transcriptions. TSAB does not handle the acquisition and transcription of spoken data itself. These tasks are handled by separate modules that are not part of the TSAB system. In our pilot installation, the speech recordings are automatically downloaded via RSS feeds of the radio stations. Then, the speech data is processed by a speech/non-speech detection module, a speaker diarization module, a speech transcription module, topic segmentation and keyphrase extraction module. Speech data and the accompanying time-aligned transcriptions are added to the TSAB internal

[1]`http://web.sls.csail.mit.edu/lectures/`
[2]`http://www.limsi.fr/tlp/audiosurf.html`

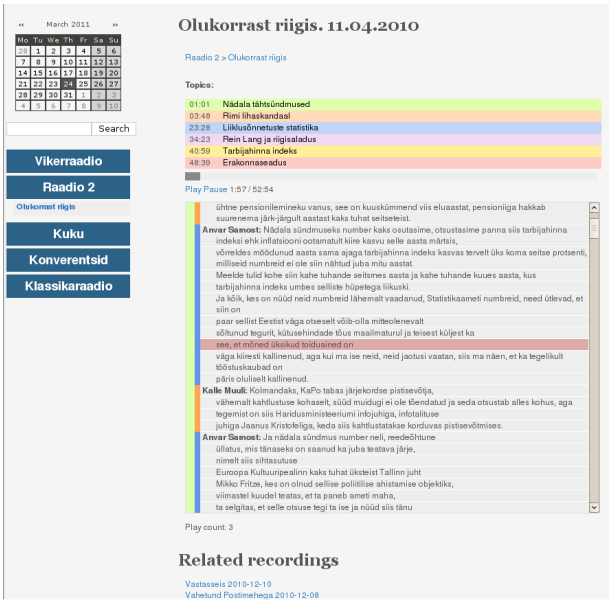

Figure 2: Example screen of the TSAB system. Shown is the view of a single recording, with the recording directory on the left, topic segmentation of the recording at the top, transcript segmented by speakers and topics in the middle, and links to automatically found semantically similar recordings near the bottom. See `http://bark.phon.ioc.ee/tsab`.

database through a scriptable command line interface. TSAB supports the Transcriber[3] XML format for speech transcriptions that enables to segment transcriptions based on speakers and topics. The segmentation markers are used to present the transcriptions in the web interface which helps to navigate in the transcription view and get an overview of the speech and discourse structure.

End-users use a web browser to navigate in the hierarchy speech recordings, search from the transcriptions, view the transcriptions and listen to the recordings. The response to a search query is a list of recordings that contain the search term(s), according to the (possibly erroneous) transcriptions. Users can instantly listen to the region of the audio file immediately surrounding the keyword hit, without bringing up the main view of the recording. The page that displays a single recording (see Figure 2) enables to get an overview of the structure of the recording, listen to the audio and view the transcript, and quickly jump to a certain point in the recording. When listening to the audio, the transcription view of the recording is synchronized with the sound stream and the current transcription segment is always highlighted. In the transcription view, segments uttered by different speakers and segments corresponding to different topics are shaded with different colors. Users can click on segments in the transcription pane to immediately start listening from any specific point. The list of topics with their descriptions (e.g., keyphrases) are also listed in front on the main transcript as a "table of contents". The page also displays links to other recordings that are similar in content.

Users need a modern web browser with a Flash plugin to use the web interface. The need for Flash plugin can be removed in the future when the HTML 5 support in browsers matures.

[3] `http://trans.sourceforfe.net`

The software is independent of the language used in the spoken audio collections. The user interface comes with English and Estonian translations.

3. Implementation

TSAB is implemented in Java and runs on a Java servlet engine. Apache Lucene is used for indexing and searching transcriptions. The most difficult part of the implementation was the fast seeking feature of the audio stream, i.e., the ability to almost instantly play any part of the recording, without the need to buffer all the audio preceding to the seek position. This turned out to be surprisingly difficult: in order easily support seeking to any position in a multimedia stream (e.g., as in YouTube) in all modern browsers, one has to do it in Flash and use a proprietary media server from Adobe, or the free Red5 media server. We wanted to avoid the need for heavy and proprietary server deployment, and implemented seeking using the following workaround: when the recording is added to the TSAB database, the audio file is split into one-minute long sections and encoded into MP3 format using a low bit rate. This enables the client implementation to seek to any position in the recording with minimal buffering, as only the section corresponding to the sought minute needs to be buffered. According to our tests, this method works well even in a low-bandwidth environment.

We use the LIUM SpkDiarization [1] toolkit for speech/non-speech detection and speaker diarization. The RWTH-ASR toolkit [2] is used speech recognition, with a three pass recognition strategy including unsupervised CMLLR and MLLR adaptation. The system has a word error rate of 28.4%. It is easy to change underlying ASR technology of the TSAB system, as long as the final ASR transcripts can be produced in Transcriber XML format.

4. Conclusions and future work

As there has been great progress in the area of spoken language processing in the last decades, it is important not to forget the importance of efficient user interfaces.

We have found that TSAB is a valuable showcase for various technologies related to speech and language processing, such as rich transcription of speech, speaker diarization, topic segmentation, keyphrase extraction and spoken term detection.

We plan to improve many aspects of TSAB in the future. One of the priorities is improving spoken term retrieval: currently, only the 1-best transcriptions are used for term detection which results in relatively low recall rates for named entities that are often incorrectly transcribed. We plan to implement more advanced spoken term detection techniques, such as subword-based stochastic matching.

5. Acknowledgements

This work was partly funded by grant #0322709s06 of the Estonian Ministry of Education and Research and by the National Programme for Estonian Language Technology.

6. References

[1] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, TX, USA, 2010.

[2] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lf, R. Schlter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Interspeech 2009*, Brighton, U.K., 2009.