AVATecH: Audio/Video Technology for Humanities Research

Sebastian Tschöpel
Daniel Schneider
Rolf Bardeli
Fraunhofer IAIS
http://www.iais.fraunhofer.de

Oliver Schreer
Stefano Masneri
Fraunhofer HHI
http://www.hhi.fraunhofer.de

Peter Wittenburg
Han Sloetjes
MPI for Psycholinguistics
http://www.mpi.nl/

Przemek Lenkiewicz
Eric Auer
MPI for Psycholinguistics
http://www.mpi.nl/

Abstract

In the AVATecH project the Max-Planck Institute for Psycholinguistics (MPI) and the Fraunhofer institutes HHI and IAIS aim to significantly speed up the process of creating annotations of audio-visual data for humanities research. For this we integrate state-of-the-art audio and video pattern recognition algorithms into the widely used ELAN annotation tool. To address the problem of heterogeneous annotation tasks and recordings we provide modular components extended by adaptation and feedback mechanisms to achieve competitive annotation quality within significantly less annotation time. Currently we are designing a large-scale end-user evaluation of the project.

1 Introduction

The AVATecH project[1] is a collaborative research project between the Max-Planck Institute for Psycholinguistics (MPI) on the one hand and the Fraunhofer Institutes HHI and IAIS on the other hand. The aim of the project is to enable researchers in the field of humanities to significantly speed up their annotation process. This process is inevitable, for example, for carrying out deep linguistic studies (Wittenburg et al., 2010; Masneri et al., 2010).

To reach this goal the Fraunhofer institutes provide audio and video pattern recognition technology for (semi-) automatic extraction of content related annotations. By integrating them into the common annotation process for linguistic research we expect a significant reduction of the overall annotation time. High potential of such technologies and tools for increasing annotation speed has been shown in (Roy and Roy, 2009).

The main challenge in the project is to tackle audio and video pattern recognition where the standard methods based on stochastic engines trained on large training sets cannot be applied to noisy field and complex lab recordings because: (1) there are only small training corpora available; (2) there are in general no models for the languages or visual setups in focus; (3) the recordings are usually of limited quality, e.g., affected by noise in the background or disadvantageous lighting conditions; (4) there are no or only few annotations that can be used to train a model. The central ideas in the AVATecH project are to adapt models to the given annotation scenario and exploit iterative feedback from the human annotator.

2 System Landscape

The system landscape of AVATecH is detailed in Figure 1. The Fraunhofer institutes are technology providers delivering recognizers in form of executables. These recognizers are integrated into existing annotation tools using a common recognizer interface that is based on a derivate of the CMDI (Component Metadata Infrastructure) specification, developed within the CLARIN research infrastructure project (Váradi et al., 2008; Broeder et al., 2010). The annotation tools are developed and maintained by the MPI. The interactive ELAN[2] tool is a widely used, open source annotation tool with a graphical frontend to annotate audiovisual content for linguistic research (Auer et al., 2010; Wittenburg et al., 2006). ELAN is not only used by many of the MPI researchers but also by a lot of other researchers worldwide. The main areas of application include language documentation,

[1] http://www.mpi.nl/avatech/

[2] http://www.lat-mpi.eu/tools/elan/

Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop, pages 86–89,
Hissar, Bulgaria, 16 September 2011.

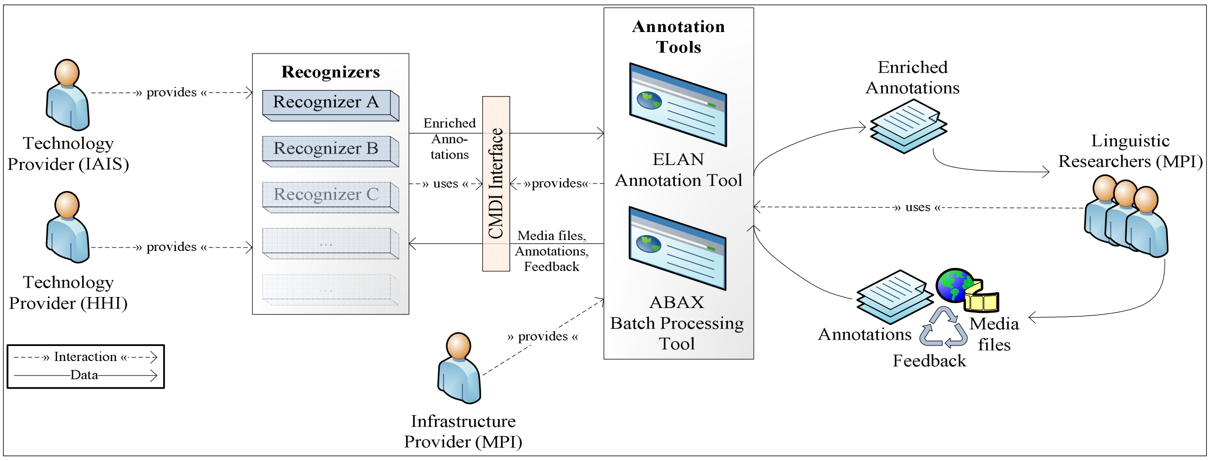

Figure 1. AVATecH System landscape

sign language research and gesture research. An additional tool, ABAX, has been created in the AVATecH project. In contrast to ELAN it is used to perform a series of annotation tasks on multiple files. ABAX provides a CMDI-interface as well. Researchers can use either ELAN or ABAX to create enriched annotations using the recognizers provided by the Fraunhofer institutes. Researchers provide media files to be annotated and, depending on the recognizer, existing annotations or additional feedback information, e.g., parameter settings, to optimize the performance of the recognizers.

3 State of work

During the beginning of the AVATecH project, we carried out intensive corpora studies to identify typical annotation scenarios and their costs in terms of hours spend by the humanities researchers. The sample corpora provided by MPI consist of 38 sub-corpora with a total file size of 730 GB and about 43,000 individual media documents. We concluded that the material is highly varying in audio and video quality (from office or lab experiments with good recording quality to field recordings in noisy environments), in language, genre (from monologues to interviews and other discourse situations), and in the amount of information that can be derived directly from the audio or video stream. This led to the conclusion that IAIS and HHI have to assemble flexible solutions in order to cope with the large variety of annotation problems.

In the first half of the project we mainly addressed three types of annotation scenarios and the utilization of user feedback.

3.1 Annotation of field recordings

Within the first scenario, researchers come back from an extensive field trip with tens of hours of unstructured media data. We aim at supporting annotation of arbitrary field recordings with only little manual interaction. The researchers provide their recordings to the analysis components via ELAN or ABAX in their usual working environment. If required, they can provide prior knowledge about the recordings, e.g., they can adjust analysis parameters or label a few segments for providing examples to a detection algorithm. After the analysis, they will obtain a pre-annotated set of field recordings, where they can quickly navigate to the portions of interest that requires more detailed manual annotation. We already integrated a number of recognizers to address this task:

Audio Recognizers

- *Audio Activity Detection*
- *Acoustic Segmentation*
- *Detection of Speech*
- *Speaker Diarization*
- *Vowel and Pitch contour detection*

Video Recognizers

- *Detection of Shot Cuts*
- *Extraction of Key frames*
- *Camera Motion / Motion Inside-the-Scene Detection*
- *Hands and Head Tracking*

3.2 Annotation of interview recordings

For a second scenario we exploit the resulting annotations of the first workflow described above. In this scenario, the researchers have a large set of interview recordings where they are just interested in the responses of the interviewee. We add further prior knowledge in the form of speech examples of the interviewer, and create separate tiers for the interviewer and the subjects of an interview situation. To address this task we incorporate widely used state-of-the-art audio analysis technology for the automatic detection of specific speakers. To use this component the researcher must provide a few minutes of samples of the desired speaker.

3.3 Annotation of sign language studio recordings

In the third scenario, the researchers want to create gesture annotations based on a corpus without any pre-annotations. The MPI-corpora contain sign language studio recordings that can be partitioned in two groups. The first group consists of single person videos, where the subject can be filmed from several positions, e.g., from above or facing the camera and at different camera distances. The other group consists of interviews with two to four people in the scene, none of them facing the camera. Resolution and quality of the recordings vary heavily depending on the sub-corpus.

Typical gesture annotations require the user to manually select the start and end point of each gesture as well as the appropriate descriptions. For each gesture, glosses for each hand can be included, as well as mouth position and information about eye aperture, gaze direction, head movements, etc. Accurate gesture analysis can be an extremely time consuming task. In the project, the videos are automatically prioritized, allowing the researchers to decide which ones are worth annotating without the necessity to view them in advance. The aim is to provide the automatic extraction of low level features (like position of the hands during a movement, average speed of the hands, duration of a gesture), allowing the researchers to focus on higher level gesture analysis. We integrated a recognizer to automatically estimate skin colour parameters and, building up on this, a recognizer for automatic hands and head detection and tracking. The latter also detects interaction between different body parts, for example, when two hands join or an arm is overlapping the face.

3.4 Using user-feedback for optimization

On the heterogeneous data in this project, some of the baseline recognizers perform poorly without additional adaptation. Moreover, in order to really speed up the annotation process, the researchers need to be able to rely on those automatic annotations that exclude data from the manual process. Hence, we investigate the potential of each analysis component to support either an adaptation mechanism or a feedback-loop mechanism or both. Furthermore, the graphical user interface will support fast correction of typical annotation errors produced by the recognizers.

By an adaptation mechanism we mean that the researchers provide examples of aspects they would like to detect, e.g., samples of a speaker for automatic speaker detection. Alternatively, they can choose from presets for typical acoustic environments, e.g., different presets for the acoustic segmentation of studio or field recordings.

By a feedback-loop mechanism we mean strategies where the user first runs a recognition process, gives feedback about the quality of the result and then runs the process with the updated information again. For example, for speaker identification the user adapts the recognizer by selecting some examples of the speaker, then runs the recognizer, and then verifies a number of segments. The recognizer uses this response to adapt the algorithm before running the process again.

To support this, the user interface must support various techniques to interact with the recognizers, e.g., to quickly jump from segment to segment, to allow a decision whether a segment is correctly labeled or not (feedback-loop), or to select segments that will be used for the adaptation mechanism.

User interaction is an essential part of our annotation workflows and it directly addresses the goal of reducing the overall annotation time since some of the originally unsupervised algorithms will not be able to provide the necessary high-quality annotation.

4 Evaluation Phase

For the evaluation, we are interested in two sets of information: qualitative statements from the human annotators that assess the annotation experience when using the AVATecH system and a quantitative evaluation of the actual annotation speed-up compared to manual annotation.

During the qualitative evaluation we are interested how the work of annotators is supported by automatic analysis, and whether annotators think that the system is beneficial. In the evaluation, subjects annotate their own material with ELAN supported by recognizers. We record their experience with a questionnaire. Within the qualitative evaluation we aim to cover all three annotation tasks mentioned above. The questionnaire consists of 20 questions addressing the quality of the ELAN interface, the productivity using recognizers during annotating, the quality of the delivered results and the overall experience. We aim to incorporate at least thirty MPI researchers or students who want to annotate their data with recognizer support.

Within quantitative evaluation, we want to measure the speed-up for annotation by using automatic tools. We ask a limited number of people to perform a specific annotation task for a subset of field recordings or studio recordings from our corpus, both with and without support from automatic analysis. The annotation will consist of labeling the speech of the interviewee, such that the researcher can quickly browse from answer to answer (for interview recordings) and labeling the gesture of the person in the video, so that the researcher can quickly see when a gesture begins, ends and what kind of motion is associated to it (for studio recordings).

To measure the annotation speed we have defined a metric that can be used not only for our evaluation, but can give a good insight about the general annotation speed of a researcher. This is not straightforward to assess, because researchers work with recordings of varying complexity (e.g. having few or many relevant events per time unit) and they are looking for different information in them, hence creating very different annotations (from very basic to multi-level, complex annotations with long descriptions).

Our general metric is based on two measures: 1) the number of created annotation blocks per unit of time; 2) the length of the media file. These values considered together will allow assessing the average annotation speed and complexity of annotation created for a given media file. Calculating these measures will be done by extending ELAN to record the overall annotation time from opening to closing the project file and to log certain annotation events, e.g., “created a new label” or “created a new segment”, with corresponding timestamps. However, if the same data is annotated twice by the same subject, the second annotation will be biased as the subject already knows the structure of the file. Therefore we penalize the automatic annotation and do it first. Moreover, we split the runs over two days (1stday: annotation with automatic tools, 2ndday manual annotation). Also we measure active vs. passive annotation time, i.e., waiting for recognizers to finish processing. This can be achieved with proper logging of the times when recognizers start and finish their execution. As population we will incorporate five to ten advanced ELAN users.

5 Future work

After carrying out the study presented above we plan to do iterations of recognizer advancements and subsequent user-reviews to further reduce the overall annotation time and to increase the user satisfaction. Also we will add more recognizers for more specialized tasks such as language-independent and -dependent forced alignment (already in an advanced stage of development), acoustic and visual query-by-example. Furthermore we expect advancements by carefully evaluating and implementing more ways of adaption and user-feedback to overcome the difficulties of heterogeneous and low-quality field recordings.

Acknowledgments

AVATecH is a joint project of Max Planck and Fraunhofer, started in 2009 and funded by both Max Planck Society and Fraunhofer Society.

References

Stefano Masneri, Oliver Schreer, Daniel Schneider, et al. 2010. *Towards semi-automatic annotations for video and audio corpora*. Proc. CSLT Workshop, LREC 2010.

Peter Wittenburg, Eric Auer, Han Sloetjes, et al. 2010. *Automatic Annotation of Media Field Recordings*. Proc. LaTECH, Workshop, ECAI 2010.

Eric Auer, Albert Russel, Han Sloetjes, et al. 2010. *ELAN as Flexible Annotation Framework for Sound and Image Processing Detectors*. Proc. LREC 2010.

Brandon C. Roy and Deb Roy. 2009. *Fast Transcription of Unstructured Audio Recordings*. Proc. Interspeech 2009.

Peter Wittenburg, Hennie Brugman, Albert Russel, et al. 2006. *ELAN: a Professional Framework for Multimodality Research*. Proc. LREC 2006

Tamás Váradi, Steven Krauwer, Peter Wittenburg, et al. 2008. *CLARIN: Common Language Resources and Technology Infrastructure*. Proc. LREC 2008

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, et al. 2010. *A Data Category Registry- and Component-based Metadata Framework*. Proc. LREC 2010