6

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek, pp. 39–45
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4515

Supporting linguistic research using generic automatic audio/video analysis

Oliver Schreer[a] and Daniel Schneider[b]

[a]*Fraunhofer Heinrich-Hertz-Institute, Berlin*
[b]*Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin*

Automatic analysis can speed up the annotation process and free up human resources, which can then be spent on theorizing instead of tedious annotation tasks. We will describe selected automatic tools that support the most time-consuming steps in annotation, such as speech and speaker segmentation, time alignment of existing transcripts, automatic scene analysis with respect to camera motion, face/person detection, and the tracking of head and hands as well as the resulting gesture analysis.

1. INTRODUCTION. Language documentation activities have produced large audio and video corpora, which can be used to investigate various topics such as the relation between spoken language and gestures or special characteristics of endangered languages. Meaningful annotations of these corpora are required as the basis for their analyses, and so the annotations ultimately contribute to the development of new theories. One of the aims of the AVATecH project (Auer et al. 2010, Tschöpel et al. 2011) is to implement algorithms that allow for automatic and semi-automatic creation of pre-annotations for such corpora, hence reducing the time needed to perform the manual annotation task. Automatic analysis of typical language documentation data (such as data from the DoBeS archive) is a difficult task due to two factors. First of all, the size of the media corpora is very significant (currently about 70 TB in the DoBeS Archive). Secondly, the recordings are highly diverse in terms of language, conditions, and situations. Effective methods for automated processing of such content are not widely available or do not exist at all.

Automatic audio/video annotation algorithms will be important for two reasons. Firstly, they lead to a dramatic decrease in the time necessary to perform simple but time-consuming pre-annotation tasks. Secondly, automation of some parts of the process can significantly increase the uniformity of the annotations created worldwide by different researchers. This would contribute to the consistency and comparability of the available language data. In this paper, we present a short overview of the tools for automatic audio/video annotation that are currently available; these have been developed within the AVATecH project. We will also present some initial results indicating the potential benefits for researchers.

2. THE AVATECH CONCEPT. The system concept of AVATecH is detailed in Figure 1. The Fraunhofer Institutes are technology providers delivering recognizers in the 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 Max-Planck-Institute for Psycholinguistics (MPI-P). The interactive ELAN tool is an open source annotation tool with a graphical interface for annotating audiovisual content for linguistic research. ELAN is now used by many different types of researchers worldwide. The main areas of application include language documentation, 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 developed by the Fraunhofer Institutes. Researchers provide media files to be annotated and specify the required parameter settings for the recognizer that will be applied. In addition, some recognizers accept existing annotations or additional feedback information, which can be used to optimize the performance of the recognizer. The main contribution of AVATecH to ELAN is that for the first time, automatic analysis tools for audio/video analysis are available and integrated into the system. In this way, the ELAN tool is not only a graphical user interface but also becomes a powerful automatic annotation engine.

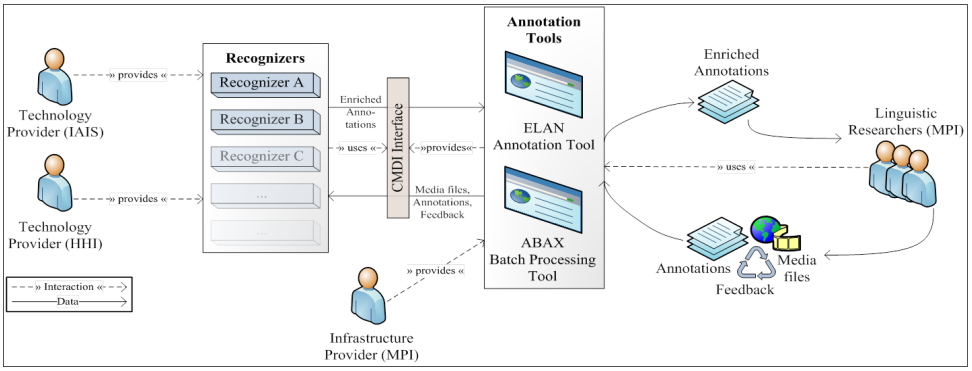

FIGURE 1: The AVATecH concept

3. AUDIO AND VIDEO ANALYSIS ALGORITHMS. All algorithms have been created with the aim of performing well on recordings with a wide variety of acoustic and light conditions as well as quite different scenarios (single vs. multiple persons). The baseline versions have also been designed to work without user interaction (except for the initial configuration by means of a few numerical parameters) to allow batch processing on multiple videos. The implementation was performed using a highly modular structure so that future automatic annotators can be easily integrated into the current framework, using as input the results provided by the previous detectors. In the following paragraphs, the recognizers which are currently available are briefly described in order to demonstrate the large

variety of analysis tools that have been developed up to now. First, we describe the audio-related detectors (from acoustic segmentation to speaker clustering), and then we introduce the video-based detectors (from shot detection to head and hands tracking).

3.1. AUDIO SEGMENTATION. For linguistic annotation, segmentation into units corresponding to utterances is of high importance but is difficult to achieve automatically without errors. This recognizer provides a fine-grained segmentation of the audio stream into homogeneous segments, e.g. between speakers or according to other significant acoustic changes such as pauses. The user can control the granularity of segmentation by tuning a corresponding feedback parameter. An important issue is the choice of the target segment granularity. Some researchers are interested in a segmentation according to who is speaking, e.g. during an interview, while others are interested in a very fine-grained utterance segmentation. The first group will perceive results from the second as being over-segmented, while the second group will consider a speaker segmentation to be under-segmented. Hence we decided to provide two baseline versions for segmentation, one focusing on speaker changes, and one optimized for fine-grained segmentation. For the utterance segmentation, we expect that more corrections will be required by the user since utterances are often not only separated acoustically but are also based on the content of the spoken words (which is not exploited by the current language-independent algorithm).

3.2. SPEECH DETECTION. This recognizer is able to label audio segments containing human speech, regardless of the language of the recording. To enhance the performance of this detector, the user can manually provide a small amount of speech and non-speech samples in order to adapt the model to the given data, which leads to a more robust detection.

3.3. SPEAKER CLUSTERING. A language-independent speaker clustering recognizer is able to find segments spoken by the same person within a given recording. The results can be used to remove utterances by the interviewer in a recording or to extract material from specific speakers from a recorded discussion. For optimization of the detection performance, we use manual user input such as the number of speakers or speaker audio samples. If the user can provide samples of a speaker, we can combine unsupervised speaker clustering with supervised speaker identification, i.e. the algorithm labels all segments where a specific known person speaks with the corresponding name.

3.4. VOWEL AND PITCH CONTOUR DETECTION. The pitch contour detector can allow researchers to graphically specify pitch contours and search for similar patterns. The detector can tag segments in audio recordings and annotate with pitch and intensity properties such as minimum, maximum, initial or final f0 frequency, or volume. The detector invokes PRAAT to calculate f0 and volume curves of the input over time. Those are then used to find characteristic segments and to annotate them.

3.5. SHOT/CUT DETECTOR AND KEY FRAMES EXTRACTOR. A shot is defined as a set of video frames that have been continuously recorded with a single camera operation and represent therefore the basic unit of a video. This recognizer is able to detect such shots and label them. All video analysis algorithms described further provide results for a given

shot and therefore rely on the results of the shot/cut detection. Sub-shots are defined as a sequence of consecutive frames showing one event, or a part thereof, taken by a single camera act in one setting with only a small change in visual content.

3.6. GLOBAL MOTION DETECTION. Accurate motion analysis allows different types of video content to be distinguished. It can be used to segment a video in order to select only those parts which are relevant for the researchers. For example, the presence of zooms and motion inside of a scene are usually the most interesting, while shots containing only panning and a low amount of internal motion are usually of little interest and can usually be discarded without further analysis. The algorithm developed performs a frame-based analysis and detects when global motion (pan, tilt, zoom in, or zoom out) occurs inside a shot. For each frame in the video, a motion vector map is computed using the Hybrid Recursive Matching (HRM) algorithm (Atzpadin et al. 2004).

3.7. SKIN COLOR ESTIMATION. A prerequisite for head and hands tracking, e.g. for gesture studies, is skin color estimation. There is no unique set of skin color parameters which can achieve good results for all recordings, and therefore typical approaches that make use of a training set to collect the parameters for skin detection on the entire dataset cannot be applied. The estimation scheme uses both the temporal information provided by the change between one frame and the next and the spatial information provided by the fact that skin color pixels tend to cluster in well defined regions. This skin color estimator does not need a training dataset but rather estimates the color ranges identifying skin color for each frame in each video.

3.8. HEAD AND HANDS TRACKING. The algorithm works first by segmenting the image in skin vs. non-skin pixels, using the information provided by the skin color estimator. The subsequent step in the detection process involves the search of seed points where the head and hands regions most likely occur. A region-growing algorithm is then applied to the seed points in order to cluster together all the skin pixels in the neighborhood. Each region is approximated by an ellipse, characterized by the position of the center, its orientation, and the length of its axes, and for tracking purposes, each of them is assigned a label (Figure 2). The tracking is performed by analyzing the change in position and orientation of the ellipses along the timeline, assigning labels based on the position of the regions in the current and previous frames.

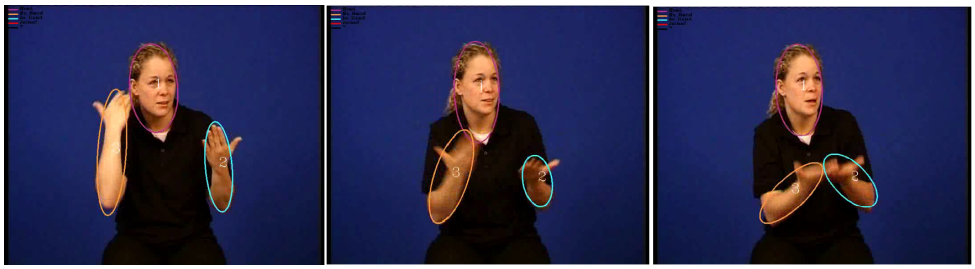

FIGURE 2: Sequence of images showing the head/hand tracking approach

4. USER INTERACTION. The expected output is very heterogeneous, and in some cases baseline recognizers can perform poorly with no additional adaptation. Furthermore, the researchers cannot accept annotation errors, such as a segment being wrongly labeled as no-speech but having speech in it (false negative). Therefore the analysis components support adaptation and feedback-loop mechanisms. The adaptation mechanism offers the researcher the opportunity to give examples of speech or video he/she is interested in, e.g. samples of a speaker for automatic speaker detection or sample segments with and without speech for the automatic detection of speech. The feedback-loop mechanism offers the user the opportunity to give feedback about the quality of the result of a first run and then perform a second process with the updated information again. For example, this could be applied for the speaker identification process: The user adapts the recognizer before running the component the first time by selecting some examples of the speaker, then runs the recognizer, and then verifies a number of segments, and the recognizer would use this response to adapt the algorithm before running the process a second time.

5. EXPERIMENTS. The development of the procedures outlined above requires intensive collaboration with researchers from the humanities in order to provide them with the most helpful and powerful tools for their annotation tasks. Therefore, a range of tests have been performed to assess how the methods can increase the effectiveness of the researchers' work. The measure of effectiveness is defined as the difference between the time necessary to create annotations for given media with and without the developed algorithms. This value cannot be easily calculated as the time necessary for annotating a time unit of media depends on factors such as 1) the purpose of the recording and contents of the media; 2) what exactly needs to be analyzed and annotated; 3) the person performing the annotation process and their expertise. Also, the level of applicability of the methods can be different for different scenarios, resulting in a different amount of help they can offer. In order to estimate the usefulness of the methods, a scenario has been created in which a researcher had to perform a number of annotation tasks aimed at different linguistic research questions. The tasks have been chosen to represent a common set of actions undertaken by a researcher annotating his/her recordings and included: 1) marking utterances of all speakers in the recording; 2) marking the size of the gesture space of a recorded person; 3) marking where speech overlaps with gestures; 4) marking specific gesturing or behavior throughout the entire recording, like nodding, raising the arms from a resting position to the level of the body, etc.; 5) marking when gesturing action happens and segmenting it into stroke, hold, and retreat. These tasks were first performed by several researchers manually, and the time necessary to carry them out was measured and averaged.

In this paper, we are describing the results of a preliminary experiment. Our aim is to gain initial insights into the potential of automatic annotation in the described scenario and use these insights to design in-depth experiments on a larger scale. For the preliminary experiment, the recognizers described above have been executed on selected exemplary recordings taken from the MPI-P archive, creating a set of initial annotations. Then, the initial automatic annotations were evaluated and manually corrected by a researcher. Figure 3 presents the time necessary to carry out three particularly time-consuming annotation tasks, namely speech utterance segmentation, gesture segmentation, and speech-gesture overlap. When the task was performed with the help of recognizers, the annotation time represents the time

required for correcting the results obtained from automatic analysis in order to make them useful for the researcher. Using the recognizers, the annotation time was greatly reduced for all three annotation tasks. The marking of utterances required some corrections of the boundaries of the annotations and also required splitting and merging some of the results of the applied algorithms. Detecting and segmenting the gestures, which are the most complex tasks, also required a significant amount of corrections. However, all test cases have proven to save a substantial amount of time required for annotation, which can be spent on other tasks.

For this preliminary experiment, recordings were not selected systematically in terms of recording conditions, length, or corresponding research scenario, but rather selected by the researchers because they were interested in the corresponding annotation task. In the next step, we plan to carry out the described experiment on a larger range of different documents and also on a larger scale such that we can analyze the effect of different recording conditions and research scenarios on the annotation speedup through automatic recognizers. High detector error rates will render our approach unusable as the number of corrections becomes intractable. Hence there is also need for investigating the minimum detector accuracy that is required for our semi-automatic approach.

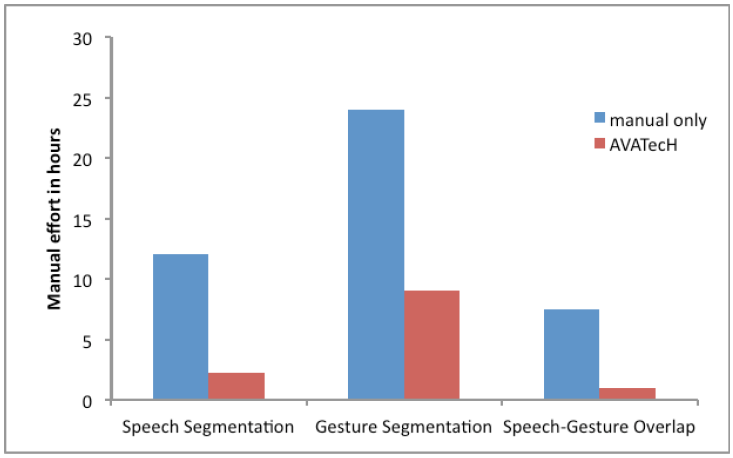

FIGURE 3: Initial evaluation results

6. CONCLUSIONS AND OUTLOOK. We have demonstrated the immense potential of automatic audio visual analysis for the processing of language documentation as a basis for linguistic research. Besides the scenarios investigated in the context of the AVATecH project, many more research domains from linguistics, and in particular gesture studies, may benefit from these techniques. An important issue for further applications of AVATecH is the availability of this technology to the international community. One challenge in this context is to find a solution as to how the technology developed by the Fraunhofer institutes can be used in the research community. Several licensing models are currently under discussion, and the common goal is to find an appropriate solution by mid 2012. It is clear that AVATecH is relevant for a large variety of research questions beyond the ones considered

in the collaborative project so far. Hence, the linguistic research community, and in particular the gesture research community, is invited to approach the AVATecH partners and to provide additional audio-visual test material and related research questions for further investigation. Only intensive collaboration between technology providers and researchers from the humanities can help to improve the methods and to adapt them to the needs and desires of the end users. We believe this collaborative work could contribute to a significant increase in the amount of annotations that can form the basis for further linguistic research.

REFERENCES

Atzpadin, Nicole, Peter Kauff & Oliver Schreer. 2004. Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *Transactions on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications* 14(3). 321–334.

Auer, Eric, Peter Wittenburg, Han Sloetjes, Oliver Schreer, Stefano Masneri, Daniel Schneider & Sebastian Tschöpel. 2010. Automatic annotation of media field recordings. In Caroline Sporleder & Kalliopi Zervanou (eds.), *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, Lisbon: University of Lisbon.

Broeder, Daan, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg & Claus Zinn. 2010. A data category registry- and component-based metadata framework. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 43–47. Valletta, Malta: European Language Resources Association (ELRA).

ELAN. Language Archiving Technology. http://www.lat-mpi.eu/tools/elan/.

PRAAT. Paul Boersma and David Weenink. Amsterdam: University of Amsterdam. http://www.fon.hum.uva.nl/praat/.

Tschöpel, Sebastian, Daniel Schneider, Rolf Bardeli, Oliver Schreer, Stefano Masneri, Peter Wittenburg, Han Sloetjes, Przemyslaw Lenkiewicz & Eric Auer. 2011. AVATecH: Audio/Video technology for humanities research. In Cristina Vertan, Milena Slavcheva, Petya Osenova & Stelios Piperidis (eds.), *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, 16 September 2011*, 86–89. Shoumen, Bulgaria: Incoma Ltd.

Váradi, Tamás, Peter Wittenburg, Steven Krauwer, Martin Wynne & Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 1244–1248. Marrakech, Morocco: European Language Resources Association (ELRA).

Oliver Schreer
oliver.schreer@hhi.fraunhofer.de

Daniel Schneider
daniel.schneider@iais.fraunhofer.de