ELAN as Flexible Annotation Framework for Sound and Image Processing Detectors

E. Auer[1], A. Russel[1], H. Sloetjes[1], P. Wittenburg[1], O. Schreer[2], S. Masnieri[2], D. Schneider[3], S. Tschöpel[3]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[2]Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany

[3]Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany

E-mail: eric.auer@mpi.nl, oliver.schreer@hhi.fraunhofer.de, daniel.schneider@iais.fraunhofer.de

Abstract

Annotation of digital recordings in humanities research still is, to a large extend, a process that is performed manually. This paper describes the first pattern recognition based software components developed in the AVATecH project and their integration in the annotation tool ELAN. AVATecH (Advancing Video/Audio Technology in Humanities Research) is a project that involves two Max Planck Institutes (Max Planck Institute for Psycholinguistics, Nijmegen, Max Planck Institute for Social Anthropology, Halle) and two Fraunhofer Institutes (Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Fraunhofer Heinrich-Hertz-Institute, Berlin) and that aims to develop and implement audio and video technology for semi-automatic annotation of heterogeneous media collections as they occur in multimedia based research. The highly diverse nature of the digital recordings stored in the archives of both Max Planck Institutes, poses a huge challenge to most of the existing pattern recognition solutions and is a motivation to make such technology available to researchers in the humanities.

1. Background

Many researchers in linguistics such as field workers and child language researchers have to work with real scenario sound and video material. Field recordings are often more challenging to process than lab recordings, for example for pattern recognition tasks. The reasons for this are manifold such as inadequate and varying position of the sensor devices (microphone, camera), various types of background noise, the need to use consumer grade devices etc. Standard speech and image recognition techniques only deliver very poor results for such recordings. Of course there are also many resources with better recording quality, but they often involve non-standard languages, long stretches of silence or regular patterns resulting from experimental settings etc. Yet, annotators would like to use any help they can get to make their work more efficient, because manual annotation is so time consuming.

There is often little knowledge about the analysed languages, so we miss formal descriptions such as proper language models. The consequence is that researchers who want to analyse this sort of material need to first carry out manual annotations based on time consuming listening and watching. In 2008 we made statistics amongst 18 teams documenting endangered languages within the DOBES[1] program to find out how much time is required for the most essential workflow steps. According to these statistics creating a transcription costs 35 times real-time, a translation into a major language 25 times real-time and for any special linguistic encoding such as morphosyntactic glossing or gesture annotation the costs in general are much higher than 100 times real time. These numbers are the reasons why an increasing number of recordings in the archives of the Max-Planck-Institutes are not annotated and even not touched any more, i.e. valuable material cannot be included in analysis of the linguistic system, theoretical considerations and cultural and cognitive studies. Advanced annotation and analysis tools such as ELAN and ANNEX [2] can facilitate the difficult work and can speed up the process slightly although no quantitative factors can be given. But these tools do not operate at the content level of the media streams.

2. Interactive Blackboard Approach

Motivated by this unsatisfying development some brainstorming between researchers and technologists of two Max Planck Institutes on the one side and sound and image processing specialists from two Fraunhofer Institutes was initiated to discuss ways out leading to a 3 year innovation project funded by MPG and FhG. Actually an old idea spelled out in the Hearsay II system (Lee et al., 1980) was brought into consideration again. In Hearsay II more or less complex independent knowledge components were operating on the speech signals each of them writing their findings on a blackboard. Other knowledge components were added that analysed the blackboard findings to finally create an automatic transcription of what was said. Such knowledge based architecture has the potential of being used to let the user interact with the low level audio and video analysis components which was one of the major requirements of the researchers at the Max Planck Institutes participating in this innovation project. In AVATecH, detector components analyze audio or video input streams and generate annotations or intermediate results. Detectors can use the output of other detectors as input, in addition to the audio and video source files.

[1] www.mpi.nl/dobes

[2] www.lat-mpi.eu/tools

After having analyzed a preliminary evaluation corpus with a variety of recordings provided by the MPIs, we found that the characteristics of the data are indeed challenging for acoustic analysis. 55 scenes from about 30 files include wind noise and similar, about 10 with reverb, about 15 with considerable background noise (engines, people, etc) and 5 with humming sounds. About 20 scenes seem to be not useful for any type of audio analysis. The speech quality itself is also varying from "indistinguishable talking" to intelligible speech. The results of acoustic segmentation, speech detection, speaker clustering and gender detection with standard algorithms optimized for broadcast data were rather disappointing as was expected. Due to the variety of languages, classic mono-lingual speech recognition could not be applied.

The initial corpus analysis resulted in a number of conclusions:

- return to the blackboard type of scenario where "detectors of various sorts" will create annotations on a new specific tier
- start experimenting with so-called low hanging fruits, i.e. simple detectors that can be integrated quickly based on existing algorithms
- have smart search and filtering methods to allow researchers to easily browse through (complex) annotation lattices
- allow the researcher to interact with the annotations and easily modify parameters controlling the functioning of the detectors so that manual tuning can be used instead of using a "one size fits all" stochastic method
- rely on existing technologies where possible with respect to the annotation and search framework and the pattern detectors

For future tests, the MPIs created a much larger collection with a few 100 gigabytes of recordings from dozens of linguistic research projects. Some of the sub-collections are also available on the public LAT archive homepage[3] and for several projects, annotations or at least transcriptions are already available. Those can be used to evaluate the results of some of the recognizers, in particular in the audio domain.

Also, while gathering the large evaluation corpus, it became evident that a component performing a time alignment of transcriptions with an audio stream would be quite useful for the linguists.

3. Annotation and Search Framework

ELAN is currently one of the most widely used media annotation tools in various linguistic sub-disciplines and beyond. It allows researchers to hook up an arbitrary number of annotation tiers referencing custom vocabularies to multiple media streams that share the same timeline. The fact that annotations cannot only be attached to a time segment but also to annotations on other tiers provides support for the creation of complex annotation structures, such as hierarchical annotations trees. The underlying EAF schema emerged from the early discussions about models such as Annotation Graph (Bird & Liberman, 2000) and it is flexible enough to cater for a large number of tiers with variable vocabularies being created by a number of (small) detectors. The screenshot in figure 1 depicts a typical ELAN window layout. ELAN has many functions including the possibility to start the well-known PRAAT [4] speech analysis software for a specific, detailed acoustic analysis.

ELAN is paralleled by the web application ANNEX that allows users to conduct analysis via the web. At present ANNEX is purely a player and data viewer, but later versions could include annotation creation and editing functions as well. Both applications are accompanied by TROVA, a flexible search engine that allows users to search for complex annotation patterns within annotation tiers, across several annotation tiers, over time and/or

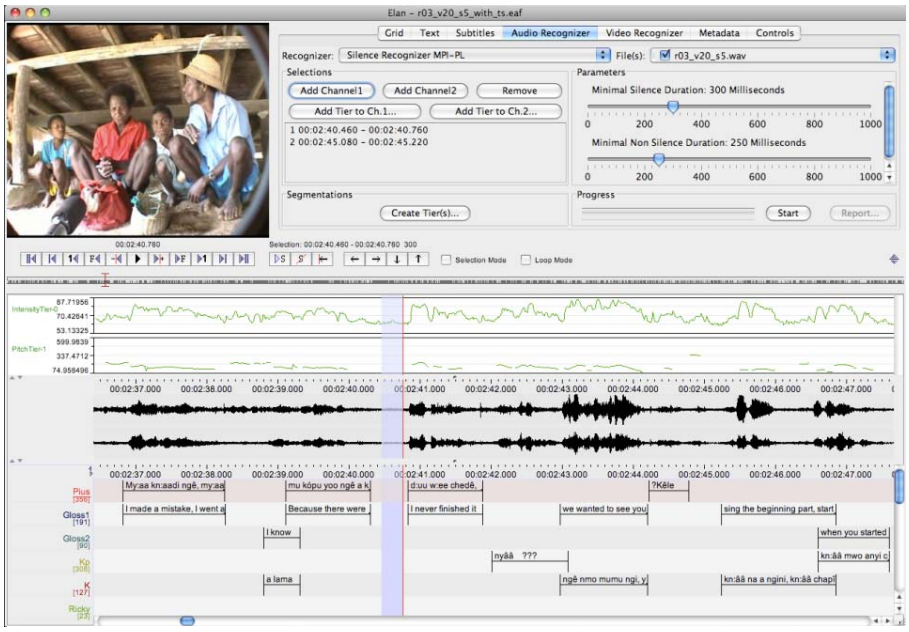

Figure 1: Use of a silence detector in ELAN 3.6.

[3] http://corpus1.mpi.nl/

[4] http://www.fon.hum.uva.nl/praat/

annotation sequences. Each pattern can be specified as a regular expression offering a large degree of flexibility. TROVA operates not only on the visualized resource, but can be used to operate on a whole selection of resources resulting from metadata searches or composed by the user. Because of indexes created at resource upload or integration time, TROVA can operate very fast on large amounts of data. The current archive contains about 60 million annotations on about 50 thousand recordings. A typical, not overly complex search operation on the entire archive displays the first results in only a few seconds.

Therefore, the current tool set consisting of ELAN, ANNEX and TROVA is an excellent starting point for improvements in the direction of adding new semi-automatic annotation and extended search functionality. In addition, the users are familiar with the user interfaces making it easy for them to adopt the new functionality.

4. First Integration Example

The first recognition component that was integrated offers simple detection of pauses (silences) in sound recordings - in fact a well-studied detection problem, the potential errors of which are known. The user graphically configures the essential parameters and receives a graphical indication of the usefulness of the choices immediately after execution. This feature of ELAN is already applied by a variety of users and it speeds up their work considerably. Some of the scenarios are:

a. In experiment result analysis, users want to quickly index or remove periods of silence in order to reduce the length of the sound wave to be analysed to a minimum.
b. Field linguists want to use the "annotation step through" function of ELAN to quickly navigate from one sequence of speech to the next, thus carrying out a first very rough selection of the material.
c. Gesture researchers can now more easily create statistics that interrelate the timing of gesture and speech segments.

It is not solely the complexity of the detection function that counts; in this particular low-hanging fruit example it is the packaging into a tool such as ELAN and the convenient graphical interaction that are attractive to researchers. The typical errors produced by such detectors are in general not dramatic, since the researchers likely use the detected segments either just for quick inspection or as a base segmentation that might be manually corrected and extended.

5. Recognizer API

Along with the first recognition component that was added to ELAN, in version 3.6.0, released in August 2008, an extension mechanism was introduced that allows producers of similar pattern recognition software components, to add their functionality to ELAN. The Recognizer API that was designed for this purpose initially offered limited functionality and the first implementation only supported audio detection components.

The second version of the API added support for video components along with more specialized data structures to be exchanged between the detectors and ELAN. At the same time, the implementation in ELAN was upgraded with support for video recognition components. To be able to provide video detectors with 2D regions of interests, preliminary graphical drawing facilities have been implemented. The user can mark e.g. rectangular areas in several video frames that circumscribe a head, face or hand or any other body part or object of interest.

The way recognition extensions are discovered and loaded has been made more versatile and robust. An extension component can now consist of multiple libraries and can be loaded without the need of adjusting the launch configuration.

A third version of the interface, published in 2010, works with I/O pipelines, XML and CSV files. This allows the creation of standalone recognizer components which can be written in any programming language and which can be accessed easily from other software than ELAN. One such tool is ABAX, the AVATecH Batch eXecutor. With ABAX, it is possible to analyze a large number of recordings. This can be done unattended, after putting all necessary recognizer parameters in a script. A wizard could be added to ELAN to make it easy to create such scripts from settings used in the current ELAN annotation session. Another advantage of the XML based interface is that it is possible to invoke recognizers running on other servers in the local network – it is sufficient to share a network drive with the server. The I/O pipeline can run over a TCP/IP connection, to send commands and receive progress and status information. Results are stored as CSV or XML files (components are free to implement either of the two) which can contain tiers (time spans with annotations) and time series (time points with measurements). It is also possible to specify auxiliary files in other formats as input or output. For example, a video analysis component could take a JPEG image as input and search time spans where the depicted object is visible or report time point / coordinate pairs of such events. Components and their parameters, input and output are specified in XML metadata files. Those 'CMDI' files can be used both for integrated and standalone recognizer components.

The registered components are presented to the ELAN user in lists, one for audio and one for video components. The user can select the recognizer of choice from the list and adjust the parameter settings. If the component does not provide a graphical user interface for setting parameter values, a standard parameter panel is fabricated for it provided the recognizer requires any parameters.

Standalone components cannot provide a parameter GUI for ELAN themselves, so they always use fabricated panels. However, they could open a GUI in a separate window if really necessary, although that is in conflict with the ability to run unattended when invoked by ABAX. When invoking components from ELAN the parameter interaction panel will be placed inside the main ELAN window. This might turn out to be insufficient for components with many configurable parameters, in which case a separate window for the panel can be made available.

We will continue to improve the user experience of working with the extension mechanism. For example, we will work on a registry where metadata about all available recognizer components is collected.

6. Low Hanging Fruit Detectors

Currently a number of such low-hanging fruit detectors are in preparation to be integrated. For audio signals we are working on the integration of

a. a noise-robust segmentation of the audio stream into homogeneous segments, which inserts boundaries e.g. between speakers or at other significant acoustic changes
b. a language-independent extraction of audio segments which contain speech
c. a simple pitch contour detector where researchers graphically specify typical contours and the detector then will look for similar patterns. An already existing AVATecH detector searches for vowels and annotates the corresponding time-spans with pitch and intensity properties such as for example minimum, maximum, initial or final f0 frequency.
d. a language-independent intra-document speaker clustering which labels identical speakers within a single document with the same ID. The latter can be used for removing the interviewer in a recording, or for extracting specific speakers from a discussion.

In the area of video detectors we are working on the integration of

a. a shot boundary detection that identifies scene changes as well as considerable changes in the video scene. Each detected scene as well as scene changes are marked by a still frame
b. a motion recognizer that detects either motion of the camera (pan, tilt) or motion in the scene
c. a face recognition detector, which identifies the number of persons in a scene
d. a detector that identifies body parts and indicates periods of movements
e. a gesture recognition tool that identifies simple hand gestures, still or moving

Currently, we are testing the behaviour of the existing detectors with respect to the variety of material we have in the archive. It is obvious that we need to study, how we can create simple to use interfaces to allow users to influence detection parameters easily and to immediately see the effects. Moreover we would like to gather feedback from users in an iterative process to improve the quality of the analysis. The TROVA search engine needs to be extended to make use of probability indications.

7. Summary

With integrating a number of detection components that create layers of annotations that can be easily used by ELAN/ANNEX and TROVA functionality, we are making a new step in facilitating the work of manual annotators. As has been seen from the very simple silence detector, which we used as first example, it can speed up the work of researchers by factors when the interaction interface is simple and when the user can stay in a well-understood tool framework. As indicted a set of first low hanging fruit detectors is being tested and will be integrated into the ELAN framework. The results will be analyzed to determine which other more complex detectors will be added and how user interaction options need to be modified to maintain attractiveness for researchers who are not per se interested in pure recognition scores but in understanding underlying mechanisms.

8. Acknowledgements

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

9. References

Bird, S., Liberman, M. (2000). A Formal Framework for Linguistic Annotation (revised version). 26 Oct 2000. http://arxiv.org/abs/cs.CL/0010033

Erman, L.D., Hayes-Roth, F., Lesser, V.R., Reddy, D. R. (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Computing Surveys (CSUR) Volume 12, Issue 2* (June 1980), Pages: 213 - 253, Year of Publication: 1980, ISSN:0360-0300

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers, 41(3), 841-849*

Crasborn, O., Sloetjes, H., Auer, E., & Wittenburg, P. (2006). Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In *C. Vetoori (Ed.), Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios (pp. 82-87). Paris: ELRA*

Auer, E., Sloetjes, H., Wittenburg, P. (2010). AVATecH Component Interface Specification Manual http://www.mpi.nl/research/research-projects/language-archiving-technology/avatech/