und mit Anmerkungen versehen von David Wirmer (Herders Bibliothek der Philosophie des Mittelalters 15). Freiburg 2008.

http://dare.uni-koeln.de/

http://wiss-ki.eu/

http://developer.berlios.de/projects/xeletor/

AV Processing in eHumanities – a paradigm shift

Wittenburg, Peter
peter.wittenburg@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lenkiewicz, Przemyslaw
przemek.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Auer, Erik
erik.auer@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Lenkiewicz, Anna
anna.lenkiewicz@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Gebre, Binyam Gebrekidan
binyamgebrekidan.gebre@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

Drude, Sebastian
sebastian.drude@mpi.nl
Max Planck Institute for Psycholinguistics, The Netherlands

1. Introduction

Speech research saw a dramatic change in paradigm in the 90-ies. While earlier the discussion was dominated by a phoneticians' approach who knew about phenomena in the speech signal, the situation completely changed after stochastic machinery such as Hidden Markov Models [1] and Artificial Neural Networks [2] had been introduced. Speech processing was now dominated by a purely mathematic approach that basically ignored all existing knowledge about the speech production process and the perception mechanisms. The key was now to construct a large enough training set that would allow identifying the many free parameters of such stochastic engines. In case that the training set is representative and the annotations of the training sets are widely 'correct' we could assume to get a satisfyingly functioning recognizer. While the success of knowledge-based systems such as Hearsay II [3] was limited, the statistically based approach led

to great improvements in recognition rates and to industrial applications.

However, most humanities and social science research does not deal with proper signals that allow to apply neither the purely statistical nor the rule-based approach. Speech is spoken in natural situations embedded in noise, it is widely spontaneous, often one has to deal with much variation for which we lack advanced models and the existing corpora are small. Thus there is no chance to apply holistic speech recognizers that take a speech signal and would produce a useful transcription.

The situation for moving image processing (video) is even worse, since we do not have an accepted target – researchers target annotations are mostly semantic functions that are associated with gestures, mimics and other body motions, which are very much dependent on cultures, situations and other parameters. Only for sign languages we can consider a situation comparable to oral speech, since we can correlate between a stream of video observations and a target transcription. However, sign languages use various information channels, which are easy to be comprehended by the human eye, but difficult to process by machinery.

This situation is not satisfying since we see that the gap between the amount of material that has been recorded by researchers and the amount of material that is available for research purposes gets larger since the available time for creating annotations did not change substantially despite new and more efficient annotation software such as ELAN [4]. Thus in psycholinguistics and in many other humanities disciplines dealing with natural scenes a new 'paradigm shift' was required.

2. Interactive Recognition Paradigm

This new paradigm is based on four equally important pillars:

- training many statistic recognizers on small phenomena and improve robustness, i.e. widely reducing the complexity of the phenomena to be recognized
- including interactive learning methods
- providing a highly efficient usability framework that brings researchers back into an active role
- make the existing and partly complex audio/video recognition technology available to the researchers

For data streams in the humanities we cannot assume that there is one recognizer that does it all. What researchers want is to have the possibility to train a recognizer to detect a specific hand movement or a specific intonation contour for example. These recognizers will then create probabilistic annotations at a specific tier. All these detectors are adding annotations ending in a complex lattice. The type of approach to realize a recognizer depends very much on the type of pattern to be detected. A cascaded recognizer, a recognizer that makes use of annotations of earlier ones, could be rule-based or based on statistics.

Interactive learning methods need to be implemented to allow the researcher to quickly improve the representation of a specific pattern including its variation. First, a single sample might be sufficient. Supervised learning techniques might help to improve the annotation accuracy to acceptable values after a few iterations.

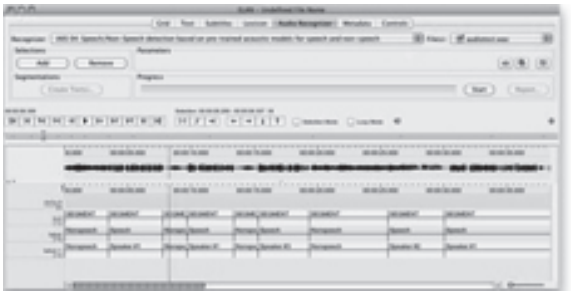

Figure 1: Three layers of annotation representing three steps of the automated analysis. First layer is the result of the uniform document segmentation, second layer is the division into speech/no-speech parts, third layer recognizes different speakers in the recording. All the steps have been performed automatically

All recognizers will create erroneous annotations, i.e. we need a framework that allows users to quickly scan annotation patterns and correct them. Here we can build on the ELAN annotation tool and TROVA search tool that have already been optimized over the years.

One of the basic assumptions of such an approach is that there are many recognizers available to the researcher. Although there exists a lot of partly complex technology in the specialist labs, it is not accessible yet. A change of culture is required to make such technology accessible, as is currently being worked out by CLARIN[1]. A standardized mechanism for invocation is required to allow starting such recognizers and to interact with them. The positive WebLicht[2] experience motivates us to continue extending the Service Oriented Architecture to audio/video services.

This new paradigm is shifting complexity to the annotation lattice. From many discussions with researchers we can expect that the method can work, since researchers are mostly not interested in 'complete' annotations, but they are interested in certain specific phenomena. For these kind of

selected annotations the interactive paradigm seems to be appropriate. In addition, we have seen in first experiments that the researchers will become engaged again, since they are not confronted with a trained machine where they do not know what the values stored in the many parameters mean. Now they can use the patterns to be looked at and the annotation created as part of their theorization process.

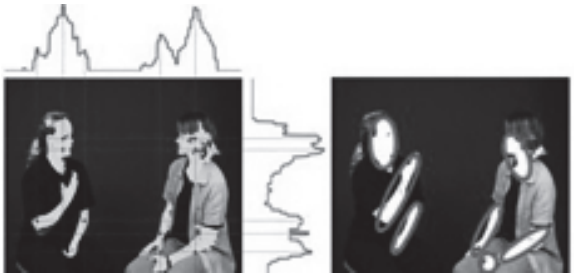

Figure 2: Left – a video frame with skin color pixels marked and histograms of those in two dimensions, X and Y. Right – example of ellipses approximating the skin areas in given image

3. The AVATecH Approach

The AVATecH project [5], started in 2009 as a joint work between MPI, IAIS[3] and HHI[4] experts, is working along the Interactive Recognition Paradigm. A number of audio/video recognizers have been implemented and integrated, the usability framework ELAN has been improved and a method for remote invocation has been established that can be extended to a Web Services scenario.

Audio Detectors

One of the recognizers provides a fine-granular segmentation of the audio stream into homogeneous segments allowing the user to control the granularity of segmentation. Another recognizer is able to label audio segments containing human speech, regardless of the language of the recording.

A language-independent speaker clustering recognizer is able to find segments spoken by the same person within a given recording (Figure 1).

A 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 for example minimum, maximum, initial or final f0 frequency, or volume.

Video Detectors

A shot and sub-shot recognizer is able to detect shots of similar video content and label them. All further algorithms rely on the results of this shot/cut detection.

Accurate motion analysis allows distinguishing between different types of video content and it can be used to segment a video in order to select only the parts, which are relevant for the researchers. For each frame in the video a motion vector map is computed using the Hybrid Recursive Matching (HRM) algorithm [6].

A skin-color detection algorithm [7] can be used to identify seed points where the hands and heads regions most likely occur. The resulting regions, where heads and hands are identified, are approximated by an ellipse for each video frame (Figure 2). After the hands and head have been labeled, the recognizer can detect strokes, gestures and relation between hands and head (Figure 3).

Figure 3: Results of the Hands and Head Tracking recognizer. On the video file the positions of hands and head are marked for every frame of the video. The annotations created are fully automated and include the time ranges in which left and right hand movement occurs, when the hands join and when there is an overlap of hand and face

User interaction

In close collaboration with the experimenting researchers the ELAN tool has been adapted to become a powerful framework to invoke the various detectors via an API described in XML and to search for annotation patterns and to manipulate them.

4. Conclusions

Automatic audio and video processing of natural scenes is a tough task and the project team from MPI, IAIS and HHI worked hard on the 4 pillars mentioned beforehand. A first number of about 10 recognizers can be used; the ELAN tool has been extended to a comfortable research environment and a mechanism supporting the required complex interaction between the ELAN (and finally the user) and the recognizers has been developed. The first experiments with researchers working on real scenes

has shown efficiency gains of about 70% only for the segmentation case which is very promising for speeding up the annotation work. In the realm of CLARIN we will collaborate with more of the technology providers in the specialist labs to tune their algorithms so that they can be integrated in this new interactive paradigm.

References

[1] **Rabiner, L., and B. Juang** (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE* 3(1): 4-16.

[2] **Hopfield, J. J.** (1988). Artificial neural networks. *Circuits and Devices Magazine, IEEE* 4(5): 3-10.

[3] **Erman, L. D., et al.** (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Comput. Surv.* 12(2): 213-253.

[4] **Wittenburg, P., et al.** (2006). Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC* 2006.)

[5] **Auer, E., P. Wittenburg, H. Sloetjes, O. Schreer, S. Masneri, D. Schneider, and S. Tschöpel** (2010). Automatic annotation of media field recordings. In *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* 2010: University of Lisbon, Portugal, pp. 31-34.

[6] **Atzpadin, N., P. Kauff, and O. Schreer** (2004). Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *Circuits and Systems for Video Technology, IEEE Transactions* 14(3): 321-334.

[7] **Terrillon, J. C., et al.** (2000). Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference.*

Notes

1. www.clarin.eu (`www.clarin.eu`)
2. weblicht.sfs.uni-tuebingen.de (`weblicht.sfs.uni-tuebingen.de`)
3. www.iais.fraunhofer.de (`www.iais.fraunhofer.de`)
4. www.hhi.fraunhofer.de (`www.hhi.fraunhofer.de`)