Tools for a combined analysis of speech and gestures

Volker Dellwo

Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn

E-mail: vde@ikp.uni-bonn.de

ABSTRACT

The availability of fast computers and recent software developments does not only enable us to carry out complex digital speech analysis procedures on an everyday basis, it also allows us to analyze visual information of the communication act by processing annotated digital video data. This study presents a selection of computer tools with which an editing and analysis process of computer readable video data for communication research can be carried out on different levels. A demonstration of the software tools will be given on an example of a project work currently carried out by the author to study the influence of moods on speech and gestures.

1 INTRODUCTION

In recent years the importance of visual information in the speech act has been paid more and more attention to. Conferences like the Orage[1] 1998 [8] and 2001 [3] show an increasing interest in a multimodal analysis (e.g. speech, gesture, prosody) from the points of view of different disciplines such as linguistics, phonetics, psychology, medicine, computer sciences, etc. The conference proceedings [3, 8] present one of the most exhaustive overviews of different current approaches in the field.

With the development of powerful computers numerous software tools have been created in different research projects in order to study speech and gestures[2], many of which are freely available for research purposes. Kipp [4 and 5] provides a good (but not exhaustive) overview of which software is available including short descriptions of each tool. Another valuable source of information on different tools is [6]. All tools currently available are annotation tools which basically provide methods for a time aligned labelling of video material.

Many researchers in the field of speech analysis may have experienced how time consuming the annotation of speech files can be; the annotation of gestures is even more time consuming. By working with my own data collection (see below) I realized that there are basically two different needs that are not served by a single software: first, the researcher needs to get a broad overview of the data that a corpus offers and second, a detailed analysis of certain passages of this corpus in regard to various speech and gesture parameters under investigation must be possible.

I want to present a set of different software tools that has been found very useful in fulfilling those needs. In the following I will give a brief description of the material under investigation followed by a description of analysis procedures that are currently carried out. As I have not compared all existing software alternatives there may be other appropriate analysis tools which could serve the here specified purposes as well. The selection described is particularly useful since it allows itself to be adapted to various different research frameworks. As research budgets are generally short another important aspect of the presented tools is that they are all available at no cost.

2 THE AUDIO/VISUAL MATERIAL

The material described here was collected through a series of interviews in Edinburgh (Edinburgh Interviews) in order to create an audio-visual corpus for mainly studying influences of emotions on speech and gesture parameters. For this purpose 18 native speakers (Ss) of English from Great Britain and the United States, all students and staff at the linguistics department of the university of Edinburgh, were interviewed in three steps: first, Ss were interviewed about their experiences with music and/or art (M&A; 5 to 10 minutes) [3]. Second, Ss were presented a one minute excerpt of a Mickey Mouse cartoon and were asked to retell it (MM; 1 to 2 minutes). Finally, Ss were interviewed about the way they perceived the terrorist attacks in the United States on September 11th 2001 (TA; 5 to 10 minutes).

The general idea behind the choice of interview topics was to trigger different emotional responses. An important fact to mention is that the interviews took place only 10 to 13 days after September 11th 2001. Opposed to the M&A interview Ss were expected to be more emotionally involved in the TA interview. It was expected that they would either show regret for the victims or fear that they themselves could be victims of such attacks. All of the Ss showed at least traces of the expected emotional involvement though some Ss remained rather neutrally formal during all three interviews.

---

[1] Orage: **Ora**lité et **Ge**stualité

[2] In the present study all visual information in the speech act is referred to as 'gestures' and all acoustic information as 'speech'.

---

[3] The time information refers to the total duration of the interview without preparation.

Audio as well as speech recordings were made with a Sony digital camcorder. In order to enhance sound quality of the speech recordings an external condenser microphone was connected to the recording device and placed about one meter away from the speaker. The material was transferred into computer readable files of the type .avi by technical staff of the media department at Jena University and was then saved on CD. This transfer of video material from tape to computer can nowadays be made on any average home computer (cf. [5] for a detailed description).

3 EDITING AND ANALYSING TASKS

Approximately 200 minutes of interview material in 54 files (3 interviews with 18 Ss) has been collected in total. A complete detailed analysis of several speech and gesture parameters may require months if not years of annotation work and in most cases this is not at all required since parts of the material only contain rudimentary gestures and are thus not valuable for an analysis. The following criteria are therefore formulated and currently carried out:

- To reduce the total amount of interview material only answers to comparable questions (e.g.: "Were you afraid when you saw the pictures [of September 11th] on television?) have to be extracted. For this purpose an appropriate tool for editing digital videos is needed (cf. 4).

- To receive an overview of the complete material, the database has to be organized with a transcription of the complete available discourse and a rough description of where in the files which gestures (e.g. iconic gestures, adaptors, etc.) are to be found. This step is henceforth referred to as the 'macro' analysis (cf. 5).

- In order to study certain passages of interest in detail, these passages have to be labelled according to certain speech and gesture parameters under investigation (e.g. phases of iconic gestures or adaptors in relation to eye lid closings, pitch movements etc. This analysis step is henceforth referred to as the 'micro' analysis (cf. 6).

In the following the software solutions to these criteria are described.

4 EDITING THE VIDEO MATERIAL

The software VirtualDub (currently version: 1.5.1) proves to be a valuable tool for editing video material. It is downloadable from the VirtualDub home page [10]. The software is a user friendly designed tool that reads video files of all common formats and allows parts of the files to be extracted to new files (these files will be in .avi format). To save disc space VirtualDub allows to use a variety of video compression algorithms. Furthermore, VirtualDub allows the sound file that accompanies a video file to be split off and saved as a separate (.wav) sound file. This is a very important feature that will be needed under 6 for a separate analysis of sound and video files.
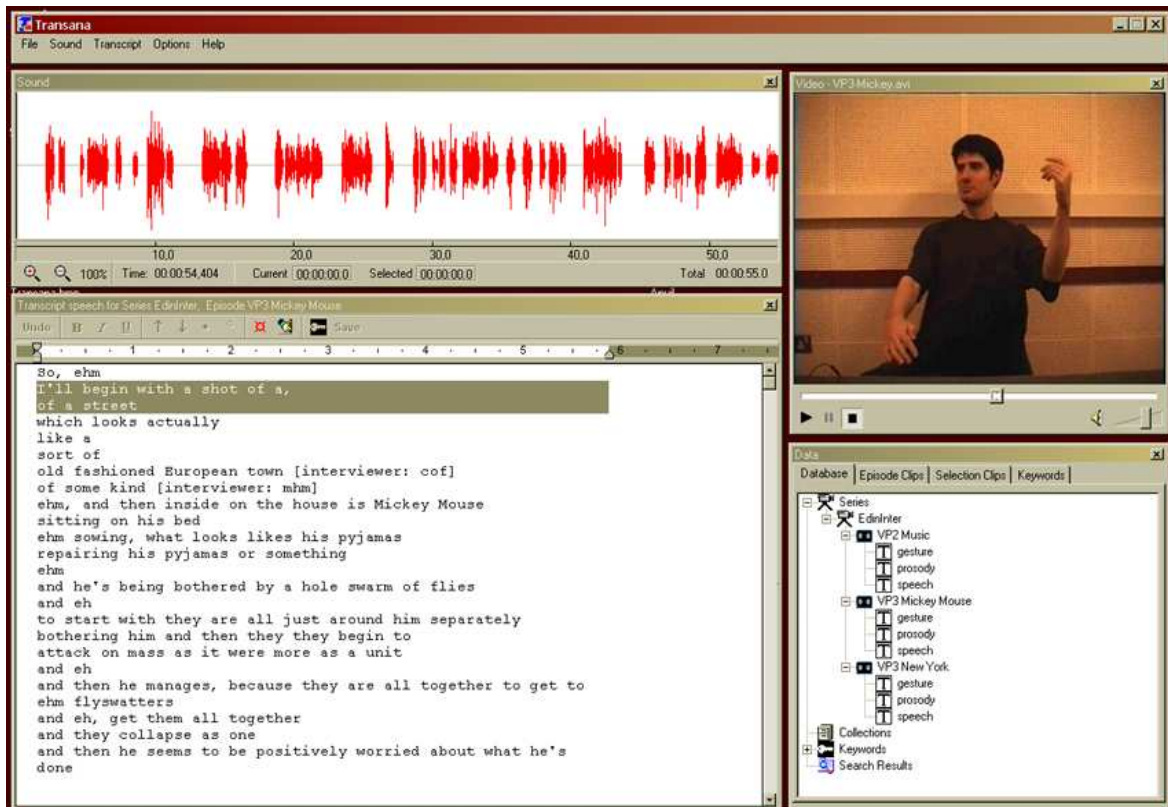

Figure 1: A screen shot of a computer session with Transana. The top window includes the program menu, the window below to the left the waveform and to the right the video. Below the waveform a literary transcription of the speech is visible, in the window below the video the database is organized.

5 MACRO ANALYSIS

Transana software is used for the macro analysis of the Edinburgh Interviews. The program is written by Chris Fassnacht in Python and supported by David Woods. It is downloadable (current version: 1.21) for nearly all versions of Microsoft Windows from the Transana home page [9]. According to the web-page, efforts are being made to provide the software for Macintosh in the near future. Transana is well documented and includes an useful electronic tutorial which can be found in the help menu. The discussion board on the Transana home page provides further valuable information (especially on supported file formats).

Figure 1 shows a screen shot of a session with Transana. The upper window includes the program menu. Below this window to the right is the video window. In the video window one of the male Ss of the Edinburgh Interviews is visible in the MM retelling task. Left of the video window is the sound window which includes a waveform of the audio data. Convenient handling mechanisms enable a quick transcription of selected aspects in the transcript window (underneath the sound window). What can be seen here is a literary transcription of the spoken discourse. An unlimited number of time anchors can be inserted in the transcription. The anchored sections will then e.g. be highlighted when the video is played back.

One of the most powerful features of Transana is its organization of data collections in the database window. The basic structure of this organization is that different series containing video file collections can be defined. All files belonging to a series are displayed including the transcriptions of these files. In the demonstration screen shot, the series is called EdinInter (Edinburgh Interviews) and it includes three videos, one for each respective interview topic. Three transcriptions belong to each video of the series: a literary transcription of the discourse, a broad transcription of prosodic features and a broad transcription of gestures. The number of addable transcriptions is unlimited. A powerful keyword notation system allows to add keywords for better accessibility of the data. Furthermore, the transcriptions are saved in Rich Text Format (.rtf) and can thus easily be edited e.g. with Microsoft Word.

Transana proves to be an extremely valuable tool for a general organization of the data of the Edinburgh Interviews. Thanks to Transana's quick file access and convenient play back methods, all videos of the collection can easily be compared and searched through by different keywords (e.g. a search for gestures of a certain type). Nevertheless, concerning a detailed analysis of gestures (e.g. gesture phases) or a comparison of different speech and/or gesture parameters, presentation modes in Transana are considered too global.

6 MICRO ANALYSIS

The micro analysis is being carried out with Anvil and Praat software:

Anvil is available from the Anvil homepage after registration with the author Michael Kipp [1]. The software
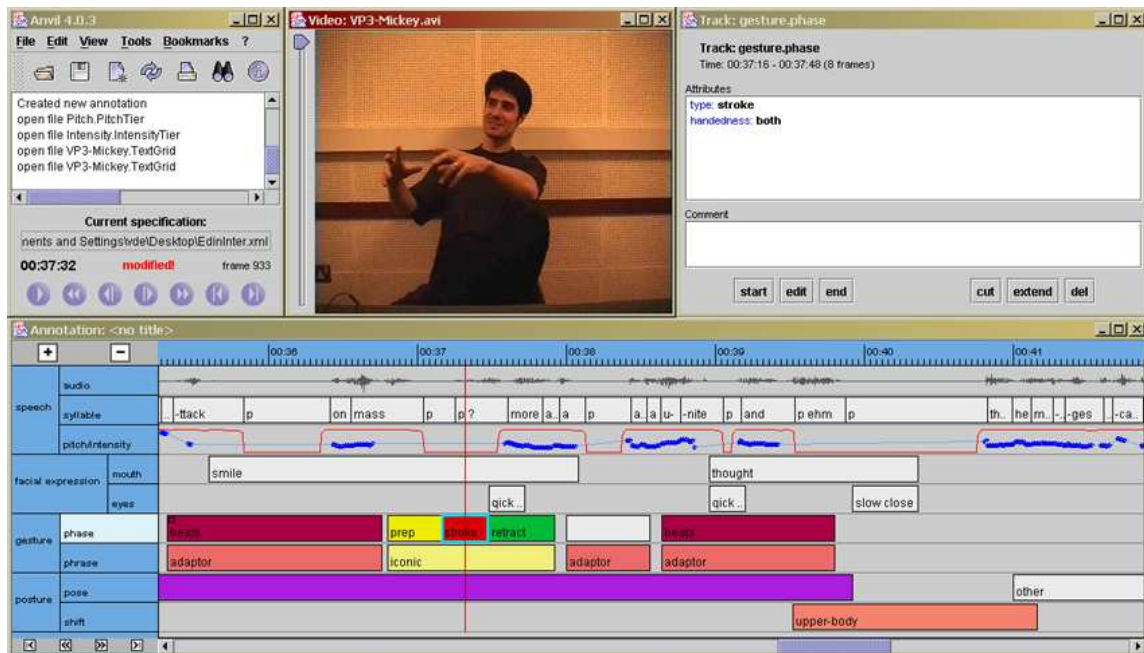

Figure 2: A screen shot of a computer session with Anvil. Four windows are visible, the Anvil menu window (top left), the video window (top mid), the track edit window (top right) and the annotation window (bottom).

is written in Java and consequently platform independent. However, a Java runtime environment as well as a Java media player have to be installed first [1]. The software is well documented [5] and supported by the author himself. According to [4], Anvil compares to many other existing tools "in that it is less ambitious" (p. 1370). This criterion is considered particularly relevant to apply Anvil to the individual research frameworks of the Edinburgh Interviews.

In Anvil, the annotations of different parameters under investigation are carried out on respective tiers which are presented in a form comparable to a musical score, that is to say a time aligned representation of all annotated parameters is visible. Figure 2 shows a screen shot of a session with Anvil. Of the three top windows, the left one includes the program menu, the centre window the video playback and the right window the track and annotation information. The large window below is the annotation window where different parameters under investigation are presented on respective tiers. In the example, tiers are summarised in four groups: speech (including a waveform, a syllable transcription, as well as pitch and intensity tier), facial expression (eye and mouth action tiers), gestures (phase and phrase tiers) and posture (pose and shift tiers)[4]. The red vertical line in the centre of the score is the cursor which currently highlights the stroke phase of an iconic gesture carried out by the S with both hands. A detailed analysis of the time organization of the different parameters can now be made: it can e.g. be seen that the retraction of the stroke phase of the iconic gesture is time aligned with a short closure of the eye lids, that on the stroke phase itself speech is paused and that the whole iconic gesture as well as the preceding adaptor are accompanied by a smile of the mouth.

A distinct characteristic of Anvil is that it allows the import of tiers from the speech analysis program Praat [2, 7]. Praat is a program for doing phonetics by computer. It is written and supported by Paul Boersma and David Weenink and offers a large variety of facilities to analyse, edit and manipulate (speech)-sound files. Praat is available from the Praat home-page [7].

The syllable tier as well as the intensity and pitch tier were created with Praat and imported into Anvil. In order to work with Praat the sound file of the video has first to be saved as a .wav file (with VirtualDub – see above). Anvil is unsuitable for speech annotation e.g. on a syllable level since resolution is based on the video file (i.e. approximately 28 pictures per second) which is too broad for syllable labelling.

Statistical analysis is further possible by exporting the annotation data from Praat and Anvil into common statistics programs like SPSS or Microsoft Excel. Though various interesting results can already be withdrawn from the Edinburgh Interviews, the low quantity of annotated data is not suitable for statistical analysis at this stage. Publications on the results are expected soon.

7 CONCLUSION

The present study has given a demonstration of how to edit, organize, and analyze a corpus of computerized video data with the tools VirtualDub, Transana, Anvil and Praat. The individual software tools are all available at no cost and run on average home computers on different platforms (VirtualDub and Transana currently only under Windows). This collection of tools proves to be a good combination to address various research frameworks as the one presented here.

ACKNOWLEDGEMENTS

REFERENCES

[1] Anvil home page: http://www.dfki.de/~kipp/Anvil

[2] P. Boersma "Praat, a system for doing phonetics by computer" *Glot International*, vol. 5(9/10), pp. 342-345, 2001.

[3] C. Cavé, I. Guaitella, S. Santi (eds.), *Oralité et Gestualité: Interactions et comportements multimodaux dans la communication*, Paris: L'Harmattan, 2001.

[4] M. Kipp, "Anvil - A generic annotation tool for multimodal dialogue", *Proceedings of Eurospeech*, pp. 1367-1370, 2001.

[5] M. Kipp "Anvil Manual". Available at the Anvil home page: http://www.dfki.de/~kipp/Anvil

[6] Linguistic Data Consortium webpage on gesture annotation tools: http://www.ldc.upenn.edu/annotation/gesture

[7] Praat home page: http://www.praat.org

[8] S. Santi, I. Guaitella, C. Cavé, G. Konopczynski, *Oralité et Gestualité: Communication multimodale, interaction*, Paris: L'Harmattan, 1998.

[9] Transana home page: http://www.transana.org

[10] VirtualDub home page: http://www.virtualdub.org

---

[4] Gesture and posture groups have been arranged following the design of a Anvil demo file by Kipp [1].