A Software Toolkit for Viewing Annotated Multimodal Data Interactively over the Web

Nick Campbell, †Akiko Tabata

Trinity College Dublin, Ireland
†Kobe University, Japan
nick@tcd.ie

Abstract

This paper describes a software toolkit for the interactive display and analysis of automatically extracted or manually derived annotation features of visual and audio data. It has been extensively tested with material collected as part of the FreeTalk Multimodal Conversation Corpus. Both the corpus and the software are available for download from sites in Europe and Japan. The corpus consists of several hours of video and audio recordings from a variety of capture devices, and includes subjective annotations of the content, along with derived data obtained from image processing. Because of the large size of the corpus, it is unrealistic to expect researchers to download all the material before deciding whether it will be useful to them in their research. We have therefore devised a means for interactive browsing of the content and for viewing at different levels of granularity. This has resulted in a simple set of tools that can be added to any website to allow similar browsing of audio-video recordings and their related data and annotations.

1. Introduction

The high degree of progress made recently in both speech recognition and speech synthesis research has unfortunately not immediately resulted in greatly improved speech interfaces for spoken dialogue systems.

This can perhaps be accounted for by the fact that whereas they model well the coding of speech into text, and vice versa, they do not yet incorporate the non-propositional content in a spoken dialogue that is an essential component of turn-management, time-management, contact-management etc., (Kendon, 1990; Bunt, 2006).

To overcome this problem, there has recently been an increase in the collection of interactive conversational speech data, often in multimodal settings, for the analysis and modelling of these additional dialogue components. This collection has resulted in an explosion of data appearing on the web (see e.g., (SSPnet, 2009; AMI, 2009; CHIL, 2009)), and consequently a need to present complex data in simple ways for fast and efficient browsing.

Perhaps the best-known example of such a large multimedia corpus is that of the European-funded AMI project (FP6-506811 (AMI, 2009)), whose Meeting Corpus was created by a 15-member multi-disciplinary consortium for research into development of a novel technology that will help groups of people to interact for business purposes.

A primary focus of AMI is on developing meeting browsers that improve work-group effectiveness by giving better access to the group's history. Another is considering how related technologies can help group members joining a meeting late or having to 'attend' from a different location. The amount of data generated by such a project is enormous and although extensive metadata annotations are available, the long download times prohibit easy access to the full corpus, which is currently being distributed physically by mail on several DVDs to interested researchers.

The present paper first briefly describes our own multimodal speech corpus and then focusses on the tools that we have designed and made available to provide easy access, before finally giving details of where the software can be downloaded for use with other similar corpora.

2. The FreeTalk Multimodal Corpus

Under a research project funded by the Grant-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science, we have been collecting and annotating naturalistic conversational speech data using both high-definition video (AVCHD, 2009), 360-degree video using a small industrial camera (a Pointgrey Flea2, (Pointgrey,) fitted with a lens taken from the SONY RPU-C251), and multichannel audio. The recordings are being made in both Japan and Ireland, to complement the Japanese JST/CREST ESP Corpus (Campbell, 2002), with English as the common language, and with occasional unrestrained inclusion of Japanese or Irish terms and phrases when they arise spontaneously in the free conversations.

The purpose of the present data collection is to study the ways in which members of a conversation indicate their participation status in a dialogue and how the flow of interaction is managed both socially and in terms of discourse control (Jokinen, 2010; Campbell, 2009). In particular, we are interested in collecting samples of non-propositional speech utterances (e.g., what Ward (Ward, 2010) has called 'grunts') and samples of gestural interaction that indicate the attentional states of the participants, to learn more about how they create shared understanding through the giving and elicitation of feedback signals.

3. Assembling Complex Data

In recording multi-party conversational interaction, we are faced with several problems with respect to viewpoint and coverage. It is possible for one camera to include all participants if they are seated around a table, however informally, and for one microphone to capture all sounds, but for optimal coverage of the interaction, it is better to make use

of multiple cameras and microphones. Having the same interaction recorded from various viewpoints allows the researcher to gain finer insights into small details of the discourse. Similarly, allocating one mic per speaker (and one for general ambiance) allows for a finer examination of any overlapping speech or of the speech of one person during the laughter of others.

We prefer to be minimally invasive in our use of technology, and not 'tether' participants or require them to monitor their placing or behaviour, though we find that most people quickly become familiar with and ignore the odd pieces of equipment that inhabit the same environment. Conversation seems to follow a "natural-order" and what might be called a "social-magnetism" soon takes over, with lively and spontaneous, even rowdy, interactions resulting when people come together for casual talk (Kendon, 1973; Jokinen & Campbell, 2008). Laughter is very common, as is overlapping speech, and cross-talk, with periods of silence appearing 'in waves' throughout the discourse.

In order to model such human interaction, we as researchers need to gain an understanding of these complex processes themselves, through observation and statistical modelling of the data. And since insight comes first through observation we need ways to quickly browse through simultaneous views (or recordings) of the same states from different viewpoints. This can be facilitated by the use of composite video montages as shown in the figures overleaf.

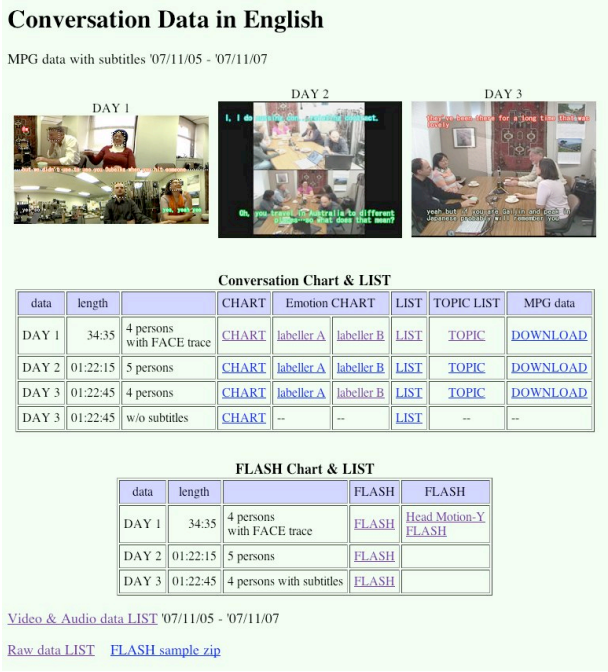
The video and audio recordings in our corpus were aligned manually, which is a time-consuming but one-off process, and the content was annotated subjectively to indicate topic, mood, and intensity of the meetings. In addition to human-produced interpretations of the speech and discourse actions, we also have machine-produced traces of movement of each participant from image-processing of the 360-degree video streams.

4. Viewing Complex Data Interactively

There are several software packages in wide general use for annotating and viewing corpus data. Prominent amongst them are Anvil (ANVIL, 2010), Elan (Elan, 2010), and Wavesurfer (Wavesurfer, 2009), and we make much use of these for the initial processing of our materials. However, while they offer pre-formatted output linking several different tiers of related information, they are not always convenient to use with very large files, and are not designed for interactive use over the internet.

Figure 1 shows the top page of the FreeTalk corpus as viewed on the Japanese server (www.speech-data.jp). It features illustrations showing samples of the recordings for each day of the Nov07 subset (Aug09 is to be added soon) and provides links to the annotations and raw video & audio sources (shown separately in figure 2).

The popular 'chart' form of viewing data is illustrated in figures 3 & 4, and forms the basis for most interactive browsing. Colour-coded representations of speech activity and participant motion allow the viewer to quickly estimate the types of social interaction. They scroll as the video plays, and subtitles (created automatically from the time information stored with the transcriptions) appear in the space between the video and the charts.



Conversation Data in English

MPG data with subtitles '07/11/05 - '07/11/07

DAY 1 DAY 2 DAY 3

Conversation Chart & LIST

data	length		CHART	Emotion CHART		LIST	TOPIC LIST	MPG data
DAY 1	34:35	4 persons with FACE trace	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD
DAY 2	01:22:15	5 persons	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD
DAY 3	01:22:45	4 persons	CHART	labeller A	labeller B	LIST	TOPIC	DOWNLOAD
DAY 3	01:22:45	w/o subtitles	CHART	--	--	LIST	--	--

FLASH Chart & LIST

data	length		FLASH	FLASH
DAY 1	34:35	4 persons with FACE trace	FLASH	Head Motion-Y FLASH
DAY 2	01:22:15	5 persons	FLASH	
DAY 3	01:22:45	4 persons with subtitles	FLASH	

Video & Audio data LIST '07/11/05 - '07/11/07

Raw data LIST FLASH sample zip

Figure 1: *showing the top page of the FreeTalk Corpus on the Japanese site. Different camera views, subtitle effects, data annotations, etc., are available.*

The 'emotion chart' plots similar bars showing speech activity, but these are colour-coded to show the perceived 'intensity' of the interactions so that a subjective estimate of the group involvement (as determined by two human annotators) may be realised (see (Campbell, 2009) and (Bunt, 2007) for further discussions of such complexity in discourse).

The traditional 'list' view is perhaps the least interesting for interactive use. It simply shows the text of the transcriptions (which we produce manually after recording) but with no visualisation of how the speech overlaps or how the speakers interact. The 'topic list' is also a manually produced text object, in the form of a spreadsheet, and contains an index for each topic item in the conversation, with time (start and end) information as well as indications of what happened at change of topic, who is mainly speaking, who is listening/reacting, and whether the mood of the scene is heated or quiet.

By switching to the exploded chart view, or 'All View', as shown in figure 4, an 'activity map' of the dialogue is displayed which allows the researcher to quickly find sections of interest in the corpus. By clicking on any of the coloured bar segments with the mouse, it is simple to switch to the relevant video scene for more detailed viewing synchronised with the derived annotations and associated data.

5. Details of the Software

The software behind this interface to the corpus has now been extensively tested and is available through the link "FLASH_sample.zip" on the top page of this site, or through the Social Signal Processing website. This compressed downloadable file contains a small sample flash video (xxx.flv), a numeric data file (xxx.dat) and a text file

Figure 2: *following the Video & Audio data link ... showing the files available for day 2 of the recordings. Composite videos have been made by combining different camera angles for easier viewing.*

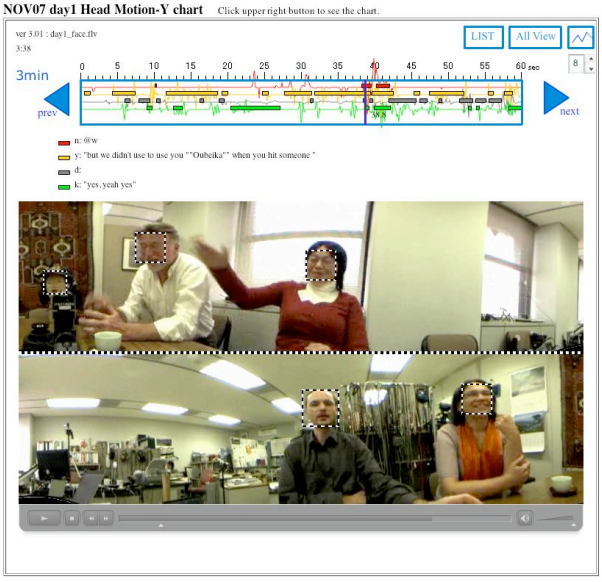


Figure 3: *interactive control of video (as a flash movie) with aligned display of speech activity (coloured bars) and movement (colour-coded data plots) overlaid.*

(xxx.txt) that are to be used as formatting examples for others who wish to view their data in the same way. It also contains a sample 'index.html' and three small programs to be included in the same directory. One, AC_RunActiveContent.is is a java script that checks for compatibility with the user's browser, and the other two are *swf* scripts, programmed in Adone's flash (Adobe, 2010) that align, format, and display the data as illustrated above. Text data, such as transcriptions, are formatted one line per utterance, with a header set of speaker's initials or other such identifiers to establish the colour code and key text (in this case, the speaker initials) to be displayed alongside the transcription as shown in figure 3. The format for data following the header is key, start time, end time, and text, with strings quoted if they include spaces. This can trivially be produced by processing the output of many transcription systems.

Figure 4: *interactive control of video (as a flash movie) with aligned display of speech activity (coloured bars) and movement (colour-coded data plots) overlaid.*

For the optional additional display of numeric data (in our case, that derived from video processing of movements (Campbell & Douxchamps, 2009)), the formatting is as shown below, here with each row containing one minute's worth of sample points for each of four speakers:

```
n
y
d
k
n 0.56 0.95 @w (laugh)
y 4.52 10.96 "hum...most of story of Manzai .. "
y 11.41 14.87 "but in Manzai  ... like"
k 14.2 14.52 hum
```

data_1: 0.01 0.02 0.02 0.01 0.01 0.02 0.01 0.00 0.02 0.01 0.01 0.00 0.79 1.00 -0.49 -0.47 0.58 0.88 0.03 -0.95 ...
data_2: -0.02 -0.00 0.12 0.28 0.37 0.70 0.17 -0.11 -0.16 -0.04 0.14 0.10 0.09 -0.06 -0.22 -0.26 -0.08 0.03 0.03 ...
data_3: -0.21 -0.18 0.09 0.20 0.02 0.15 0.06 0.07 -0.29 0.01 -0.14 0.21 -1.09 0.40 0.27 -0.16 -0.14 0.07 -0.91 ...
data_4: 0.70 0.32 0.15 -0.02 -0.23 -0.02 0.32 0.14 0.02 0.11 0.05 0.06 0.13 0.00 -0.17 -0.16 0.12 0.14 -0.01 -0.04 ...

To use these programmes with novel content, it is necessary to convert any video recordings to the flash movie format using proprietary software supplied by Adobe (Adobe, 2010), and then to ensure that the accompanying data files are formatted according to the samples included (above).

6. Downloading and Use

The software is essentially free, and free of conditions, but does require access to the Adobe flash maker software, a commercial product, (see (Adobe, 2010)) though a Google search brings up many open-source or free-download equivalents which we have not tested.

The corpus is distributed under a Creative Commons Non-Commercial Attributive license (License,) whereby we expect researchers who have added further levels of annotation to make those annotations (or corrections to the original pre-existing annotations) available to us and by extension to the wider research community.
The interactive pages are available for viewing at the main site: http://www.speech-data.jp/ (see 'TableTalk' for the English multimodal data and 'Conversations' for Japanese telephone conversations) and we are grateful to the Social Signal Processing Network (SSPnet, 2009) for providing space on their European server for faster download of raw data files: http://freetalk-db.sspnet.eu/ (FreeTalk, 2009). Passwords can be obtained from *nick@tcd.ie* if required. We would of course be very grateful for an acknowledgement if these tools are found to be useful.

7. Summary & Conclusion

Viewing large amounts of corpus data can be very difficult, particularly when streaming it over the web. Our work has resulted in a method for presenting audio-visual materials in a compact form that allows rapid search and retrieval while also at the same time offering a compact and informative view of discourse moved and actions in a dialogue.
The display engine takes audio or video data that has accompanying transcriptions or annotations and provides a simple way to view and download selected portions interactively from a project web-page.
This paper has also presented a brief overview of the FreeTalk corpus and has described the set of software that we have developed to browse it. This software is hereby placed in the public domain and can be downloaded for research use from the web addresses given above.

8. Acknowledgements

Much of this work was carried out in Japan while the first author was employed by the National Institute of Information and Communications Technology (NiCT), and by the Advanced Telecommunications Research Institute (ATR) in Kyoto, with partial support from the Japanese Government 'kaken' funding of the JSPS and the Monbu Kagakusho [Ministry of Education, Culture, Sports, Science and Technology (MEXT)]. It is being continued at NAIST and at Trinity College, the University of Dublin, with grateful thanks to the Science Foundation Ireland.

9. References

Kendon, Adam, (1990) *Conducting Interaction: Patterns of Behaviour in Focused Encounters*. Cambridge: CUP.

Bunt, H., Dimensions in Dialogue Act Annotation. Proceedings LREC 2006, Genova.

SSPNet is a European Network of Excellence fostering and supporting research activities in Social Signal Processing, the new, emerging domain aimed at bringing Social Intelligence in computers. http://sspnet.eu/

Augmented Multi-party Interaction (http://www.amiproject.org)

CHIL: Computers in the Human Interaction Loop (http://chil.server.de/)

AVCHD: new high definition (HD) digital video camera recorder format recording 1080i*1 and 720p*2 signals by efficient codec technologies http://www.avchd-info.org/

Pointgrey cameras http://www.ptgrey.com/products/flea2/

Campbell, N., "Recording Techniques for Capturing Natural Every-Day Speech", LREC2002, May 2002

Jokinen, K. (forthcoming). Gesture Activity and the Synchrony of Communication.

Campbell, N., "An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data". Proc INterspeech 2009.

Ward, N., "Non-Lexical Conversational Sounds in American English", to appear in Pragmatics and Cognition (forthcoming)

Bunt, H., Multifunctionality and Multidimensional Dialogue Act Annotation. In: E. Ahlsen et al. (ed.) Communication - Action - Meaning, A Festschrift to Jens Allwood. Gothenburg University Press, August 2007, pp. 237 259, 2007.

Kendon, Adam and Andrew Ferber (1973) "A description of some human greetings". In R. P. Michael and J. H. Crook (eds.) *Comparative Ecology and the Behaviour of Primates*. London: Academic Press. 591668.

Jokinen, K. and N. Campbell (2008). "Non-verbal Information Sources for Constructive Dialogue Management". LREC-2008. Marrakesh, Morocco.

ANVIL; The video annotation research tool: http://www.anvil-software.de/

Elan is a tool for the creation of complex annotations on video and audio resources: http://www.lat-mpi.eu/tools/elan/

WaveSurfer is an Open Source tool for sound (and video) visualization and manipulation: http://www.speech.kth.se/wavesurfer/

Flash: http://www.adobe.com/products/flash/

Nick Campbell, Damien Douxchamps (2007) "Robust real time face tracking for the analysis of human behavior", pp.1-15, in *Machine Learning & Multimodal Interaction*, Springers LNCS series, 4892.

Creative Commons; http://creativecommons.org/about/licenses/

The FreeTalk Corpus is available for download from the SSPNet : http://sspnet.eu/2010/02/freetalk/