A Software Toolkit for Viewing Annotated Multimodal Data Interactively over the Web
((- abstract for review -))

Nick Campbell & Akiko Tabata

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

Abstract

This paper describes a software toolkit for the interactive display and analysis of automatically extracted annotation features of visual and audio data collected as part of a multimodal spoken dialogue corpus. The FreeTalk Multimodal Conversation Corpus is freely available for download from sites in Europe and Japan but consists of several hours of video and audio recordings from a variety of devices as well as subjective annotations of the content and derived data from image processing. It is unrealistic to expect researchers to download all of the corpus before deciding whether it will be useful to them in their research, so we have devised a means for interactive browsing of the content and 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.

Index Terms: multimodal conversation corpora, discourse features, interactive display, image and audio data, conversation analysis

1. Introduction

The high degree of progress made 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 extra-propositional content in a spoken dialogue that is an essential component for turn-management, time-management, and contact-management etc., [1, 2]. 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., [3, 4, 5]), and consequently a need to present complex data in simple ways for fast and efficient browsing.

Perhaps the best-known example of a large multimedia corpus is that of the European-funded AMI project (FP6-506811 [4]), whose Meeting Corpus is created by a 15-member multi-disciplinary consortium dedicated to the research and development of technology that will help groups interact better.

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 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 grant funded by the Grant-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science, we have been collecting and annotating conversational speech data using both high-definition video (AVCHD, [6]), 360-degree video using a small industrial camera (Pointgrey Flea2, [7]), and audio. The recordings are being made in both Japan and Ireland, with English as the common language, with occasional inclusion of Japanese or Irish terms and phrases when they arise spontaneously in the free conversations.

The purpose of this data collection is to study the ways in which participants indicate their participation status in a conversation and how the flow of interactions is managed both socially and in terms of discourse control [8, 9]. In particular, we are interested in collecting samples of non-verbal speech and gestural interaction that indicate the attentional states of the participants, and how they create shared understanding by giving and eliciting feedback signals.

The video and audio recordings in our corpus are first aligned manually, which is a time-consuming but one-off process, and the content is annotated subjectively. 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.

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 try not to be invasive in our use of the technology, and will not tether participants or require them to monitor their placing or behaviour, and 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 to talk [11, 12]. 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 need first to gain an understanding of these processes themselves, through observation and statistical modelling of the data. And since insight comes first through observation we need ways to observe simultaneous views (or recordings) of the same states from different viewpoints. This is achieved by the use of composite video montages as shown in the figures.

4. Viewing Complex Data Interactively

Figure 1 shows the top page for the corpus when viewed on the Japanese server (www.speech-data.jp). It includes illustrative images showing samples of the recordings for each day of the Nov07 session (Aug09 is to be added soon) and links to annotations and raw video & audio sources (shown separately in figure 2). The 'chart' view is illustrated in figures 3 & 4, and forms the basis for most of the interaction. Here, colour-coded representations of speech activity and participant motion 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.

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 [9, 10] for further discussion of complexity in discourse).

The '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.

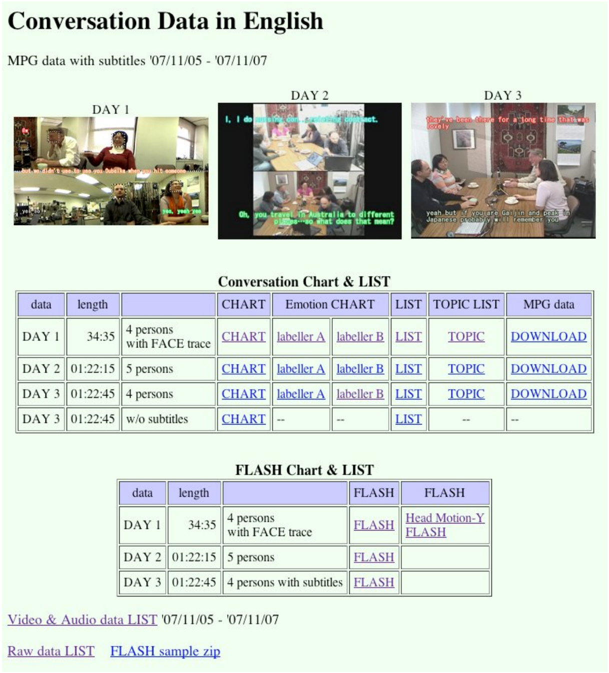
By switching to the exploded chart view, or 'All View', as shown in figure 4, an 'activity map' of the dialogue is displayed which allows the researcher to quickly find sections of interest in the corpus. By clicking on these, it is easy to switch to the relevant video scene for more detailed viewing 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 (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 programs to be included in the same directory as the html index file. 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 [13] that align, format, and display the data as illustrated above.

To use these programmes with novel content, it is necessary first to convert any video recordings to a flash movie format using proprietary software supplied by Adobe [13], and then to ensure that the accompanying data files are formatted according to the samples included.

Text data, such as transcriptions, should be formatted one line per utterance, with a header set of speaker's initials or other such identifying codes to establish the colour code and identifiers (in this case, speaker initials)



Conversation Data in English

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

DAY 1
DAY 2
DAY 3
but we didn't use to use you Oubeika when you hit someone
yes, yeah yes
Oh, you travel in Australia to different places...so what does that mean?
yeah but if you are Gaijin and peak in Japanese probably will remember you

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

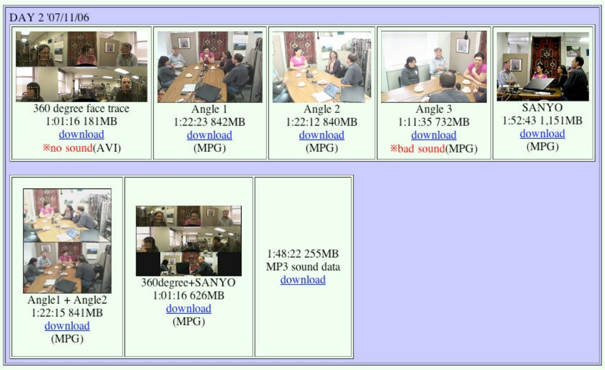
Figure 1: *showing the top page of the FreeTalk Corpus on the Japanese site. Different camera views, subtitle effects, data annotations, etc., are available.*

to be displayed alongside the text as shown in figure 3. The format for data after the header is id, start time, end time, and text, with strings quoted if they include spaces:

```
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
y 15.78 16.79 hum...
n 24.9 25.22 @w (laugh)
```

For the optional additional display of numeric data (in our case, that derived from video processing of movements [14]), the formatting is as shown below, here with each row containing one minute's worth of sample points for each of three speakers:

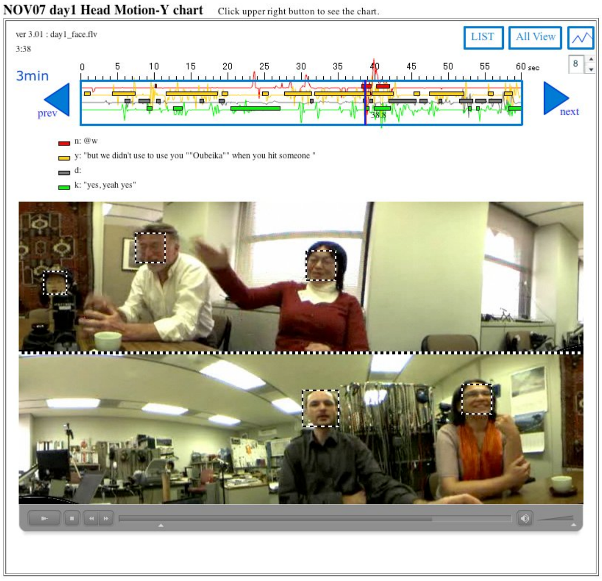
data_1: 0.01 0.02 0.02 0.01 0.01 0.02 0.01 0.00 0.02 0.01 0.01 0.00 0.79 1.00 -0.49 -0.47 0.58 0.88 0.03 -0.95 . . .
data_2: -0.02 -0.00 0.12 0.28 0.37 0.70 0.17 -0.11 -0.16 -0.04 0.14 0.10 0.09 -0.06 -0.22 -0.26 -0.08 0.03 0.03 . . .
data_3: -0.21 -0.18 0.09 0.20 0.02 0.15 0.06 0.07 -0.29 0.01 -0.14 0.21 -1.09 0.40 0.27 -0.16 -0.14 0.07 -0.91 . . .
1
data_1: -0.01 -0.00 -0.01 -0.01 -0.01 -0.01 -0.00 0.00 -0.00 -0.01 -0.01 0.01 -0.02 -0.01 -0.02 -0.01 0.01 0.77 . . .
data_2: -0.07 -0.16 -0.03 0.06 0.03 -0.02 -0.00 0.02 -0.03 0.01 0.01 -0.01 0.00 0.02 -0.01 -0.01 0.01 0.02 0.02 . . .



DAY 2 '07/11/06

360 degree face trace 1:01:16 181MB download ※no sound(AVI)	Angle 1 1:22:23 842MB download (MPG)	Angle 2 1:22:12 840MB download (MPG)	Angle 3 1:11:35 732MB download ※bad sound(MPG)	SANYO 1:52:43 1,151MB download (MPG)
Angle1 + Angle2 1:22:15 841MB download (MPG)	360degree+SANYO 1:01:16 626MB download (MPG)	1:48:22 255MB MP3 sound data download		

Figure 2: *following the Video & Audio data link ... showing the files available for day 2 of the recordings.*



NOV07 day1 Head Motion-Y chart Click upper right button to see the chart.

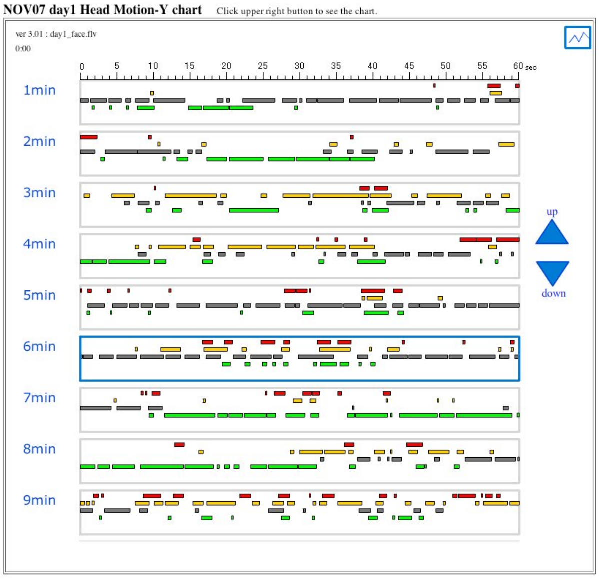
ver 3.01 : day1_face.flv
3:38

🟥 n: @w
🟨 y: "but we didn't use to use you ""Oubeika"" when you hit someone "
⬛ d:
🟩 k: "yes, yeah yes"

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.*

data_3: 0.04 0.01 0.02 0.02 0.02 -0.01 0.00 -0.01 -0.01 -0.00 -0.01 0.01 -0.00 0.00 0.02 0.00 0.00 0.01 0.02 0.03 . . .
2
data_1: -0.01 0.00 -0.02 0.01 -0.02 -0.02 0.00 -0.01 0.00 -0.02 -0.01 -0.00 -0.00 -0.01 0.00 -0.00 -0.01 -0.01 . . .
data_2: 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 . . .
data_3: 0.03 -0.01 -0.24 -0.13 -0.01 0.10 0.05 0.06 0.09 . . .

In the data sample given above, the intervening integer numbers simply serve as comment lines and are ignored in the display processing.



NOV07 day1 Head Motion-Y chart Click upper right button to see the chart.

ver 3.01 : day1_face.flv

0:00

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.*

6. Downloading and Use

The software is essentially free, and free of conditions, but does require access to the Adobe flash maker software (available from [13]), a commercial product, although a Google search appears to bring up many open-source or free-download equivalents which we have not tested.

The corpus is distributed under a Creative Commons Attributive license [15] 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 from http://www.speech-data.jp/ (see nov07 for the multimodal data and esp_c for Japanese telephone conversations) and we are very grateful to the Social Signal Processing Network [3] for providing space on their European server for faster download of raw data files: http://freetalk-db.sspnet.eu/.

We would of course be very grateful for an acknowledgement if these tools are found to be useful.

7. Summary & Conclusion

This paper has presented a brief overview of the FreeTalk corpus and has described software that we have developed to browse it. This software is hereby placed in the public domain and can be downloaded 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. It is being continued at NAIST and at Dublin University, Trinity College (with grateful thanks to the Science Foundation Ireland, SFI) and with partial support from the NiCT through the Japanese Government 'kaken' funding of the JSPS and the Monbu Kagakusho [Ministry of Education, Culture, Sports, Science and Technology (MEXT)]. Both authors would like to offer particular thanks to professor Toshiyuki Sadanobu of Kobe University for his continuing cooperation and advice with the data collection and analysis.

9. References

[1] Kendon, Adam, (1990) *Conducting Interaction: Patterns of Behaviour in Focused Encounters*. Cambridge: Cambridge University Press.

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

[3] 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/

[4] AMI: Augmented Multi-party Interaction (http://www.amiproject.org)

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

[6] 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/

[7] Pointgrey cameras http://www.ptgrey.com/products/flea2/index.asp

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

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

[10] 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.

[11] 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.

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

[13] Flash: http://www.adobe.com/products/flash/

[14] 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.

[15] http://creativecommons.org/about/licenses/