Tools & Resources for Visualising Conversational-Speech Interaction

1 Introduction

With ever-growing increases in the amount of data available for speech technology research, it is now becoming increasingly difficult for any one individual to be personally familiar with *all* of the data in any given corpus. Yet without the insights provided by first-hand inspection of the types and variety of speech material being collected, it is difficult to ensure that appropriate models and features will be used in any subsequent processing of the speech data.

For data-handling institutions such as ELDA (the European Evaluations and Language-resources Distribution Agency [1]) and LDC (the US Linguistic Data Consortium [2]) whose main role is the collection and distribution of large volumes of speech data, there is little need for any single staff member to become familiar with the stylistic contents of any individual corpus, so long as teams of people have worked on the data to verify its quality and validate it as a reliable corpus. However, for researchers using that data as a resource to help build speech processing systems and interfaces, there is a good case to be made for those individuals to become familiar with the contents and characteristics of the speech data in the corpora that they use.

It is perhaps not necessary (and often physically very diffifficult) to listen to all of the speech in a given corpus but it is essential to be able to select in a non-random manner specific sections of the corpus for closer inspection and analysis. If the data is transcribed, the transcriptions will provide the first key into the speech data but there are many aspects of a spoken message that are not well described by a plain text rendering of the linguistic content. Matters relating to prosody, interpretation, speaking-style, speaker affect, personality and interpersonal stance are very difficult to infer from text alone [3], and almost impossible to search for without specific and expensive further annotation of the transcription.

In Japan, at ATR [4], we have now collected several thousand hours of conversational speech data and have produced a web-based interface with cgi-scripts programmed in Perl that incorporate Java and JavaScript to facilitate first-hand browsing of the corpora. Some of the features of this software will be described in the sections below. In the following text, Section 2 illustrates the top-level interface to the data, Section 3 gives an example of an interface that offers fast browsing based on dialogue structure, and Section 4 illustrates facilities for the display and retrieval of multi-modal data

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	

1min flash movies

Day 3 Head motion X graph
Day 3 Head motion Y graph
Day 3 Conversation Chart

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

Fig. 1. The top page for accessing ATR dialogue data at http://feast.atr.jp/non-verbal/project/html_files/taba/nov07/ showing some of the annotated files for a series of video conversations (password & userid are available from *nick@tcd.ie*)

2 Browser Technologies

With the growing recent interest in processing multimodal interaction, beginning with projects such as NIST Rich Transcription [5], AMI [6], and CHIL [7], there has been considerable research into collecting and annotating very large corpora of audio and visual information related to human spoken interactions [8], and subsequently huge efforts into mining information from the resulting data [9] and making the information available to researchers in various related disciplines [10]. Consequently, much research has also been devoted to interface and access technologies, particularly using web browsers [11]. Our own corpora illustrate

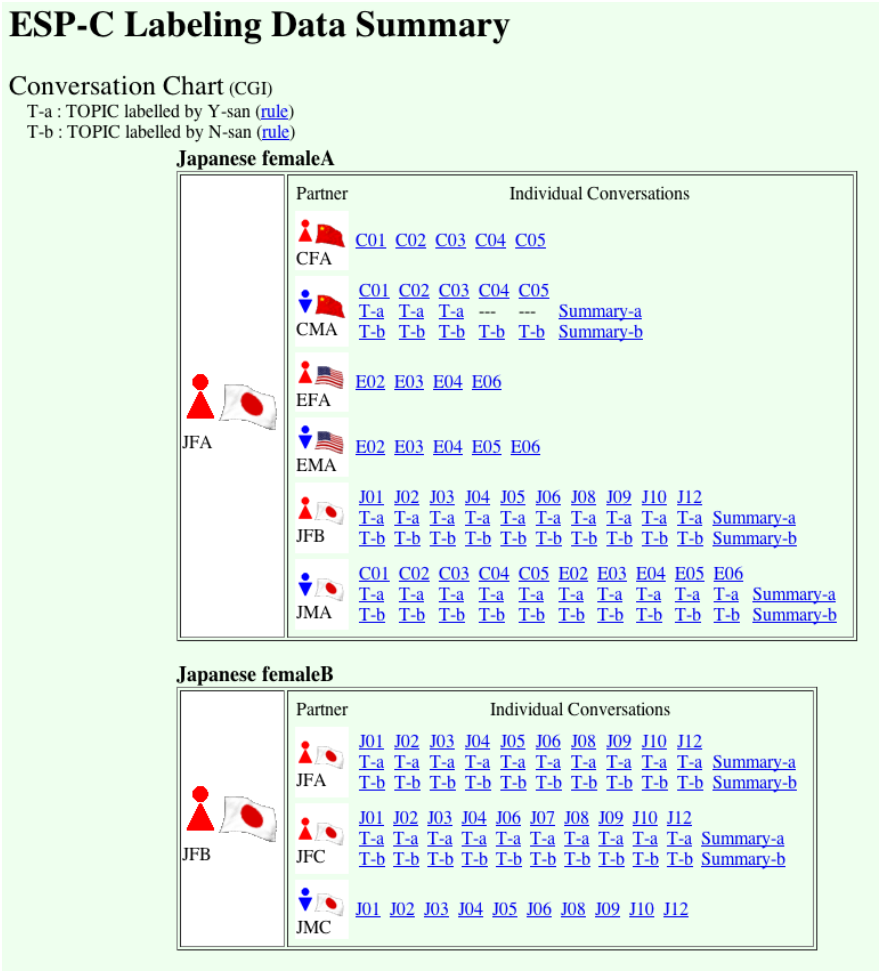

Fig. 2. The top page for accessing Japanese telephone conversations at http://feast.atr.jp/non-verbal/project/html_files/taba/top_esp-c/ showing interlocutor (by nationality and sex) and topic number, with access to summary data and topic annotations as well as to the actual sequential conversations.

different forms of spoken dialogue and are related by contextual features such as participant identity, nature of the interlocutor, mode of conversation, formality of the discourse, etc. They are stored as speech wave files with time-aligned transcriptions and annotations in the form of equivalently-named text files. Since they come from various sources, there is no constraint on file naming conventions so long as there is no duplication of identifiers. The files are physically related by directory structure and can be accessed through a web-page which hides the physical locations and provides access information in human-readable form.

Examples are given in Figures 1 & 2 which show the top-level pages for two sections of the corpus. The pages provide access to all the conversations from each participant, grouped in one case (Fig. 1) according to serial order of the dialogue sequence, and another (Fig. 2) by interlocutor, showing topic of the conversations as determined by manual labelling of the speech.

These summary pages allow the browsing researcher to visualise the style and content of each corpus with only minimal effort. In each case, the LIST option offers a quick route to the full transcriptions, although these themselves may be long and difficult to visualise.

Because of the very large number of conversations in the ESP_C subset, clicking on any link (shown in Figure 2) will bring up a text-free page, such as the one illustrated in Figure 3, which facilitates direct browsing access to the data as well as allowing immediate visualisation of its structure. We have found that this graphical way of illustrating a conversation speeds up access by allowing the researcher to see the dynamics of the discourse before having to access its transcription as text. The display makes use of the time-alignment information inherent in each transcription.

3 Browsing based on Dialogue Structure

Whereas complete manual transcriptions are available for most conversations in the corpus, the difficulty of time-aligning such texts for graphic display is well known to conversation analysts, who have devised orthographic layout conventions that allow visualisation (to some extent) of the timing and sequential information of the dialogues [12, 13]. Since overlapping speech is common in conversational interactions, the nature of the discourse can often be estimated from the structure of these overlaps.

Figure 3 demonstrates one solution to the problem of displaying and time-aligning such speech data. We took advantage of the graphical interface of an interactive web page to plot colour-coded utterance sequences for maximal visual impact. This screenshot shows two speakers' time-aligned utterance activity. Here, each speaker is shown in a different colour (pink and blue for the two speakers in this case) and each utterance is accessible by mouse-based interaction. Moving the mouse over a bar reveals its text (see e.g., the last row in the figure) and clicking on the bar will play the speech-wave associated with the utterance. This graphical form of layout makes it particularly easy to search utterance sequences based on dialogue structure and speech overlaps.

A more conventional view of the transcriptions can be accessed by clicking the LIST option in the upper right-hand corner of this page. This reveals text in the form shown in Figure 4, with utterance timing, speaker, and transcription displayed vertically. Two modes of audio output are offered for dialogue speech, since it is sometimes preferable to hear a stereo recording, which provides simultaneous natural access to both speaker's overlapping segments, and sometimes better to hear a monaural recording, where overlapping speech does not intrude on ease of listening. Separate speech files are employed in each case.

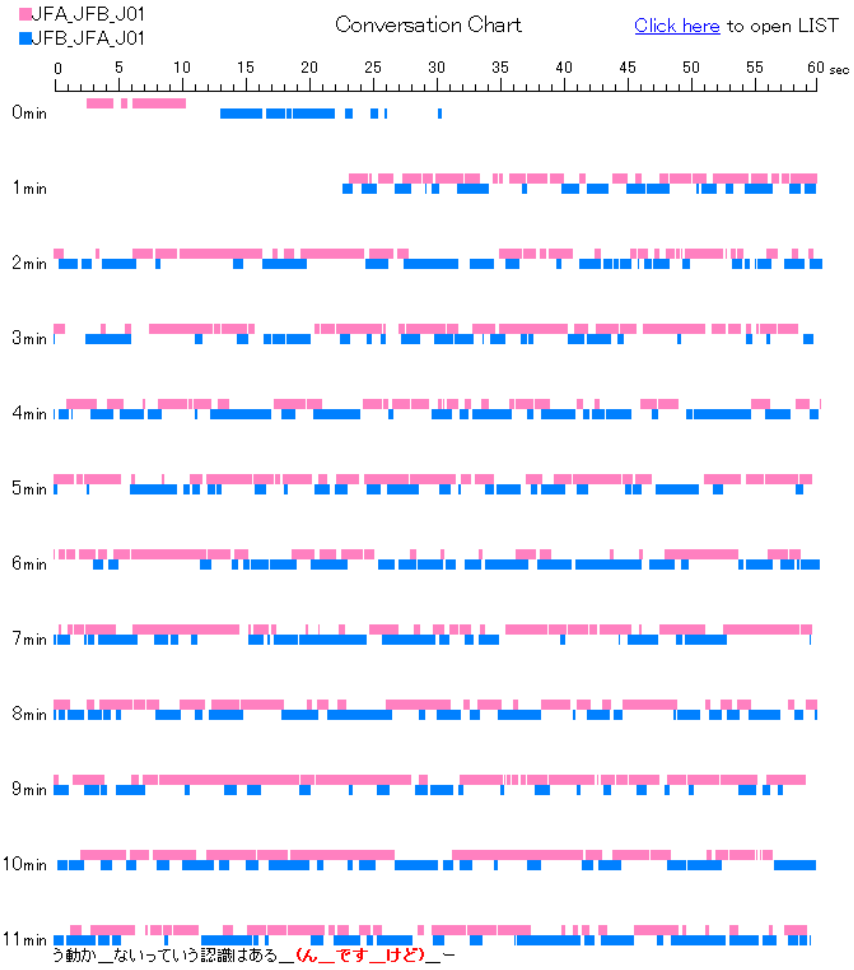

Fig. 3. A page showing speech interaction in a Japanese telephone conversation accessed from Figure 2, where one speaker is illustrated in pink, and the other in blue, allowing immediate visualisation of the overlaps and the dynamic structure of the interaction. Mousing over a bar displays its text (shown here in Japanese) and clicking on it brings up the speech data in a separate window for browsing

Search is an essential facility for any corpus, and several ways are offered for thus constraining the displayed data to specific subsets. A fast Google-type search facility was reported in [14] based on the Swish-E public-domain search-engine [15] and using text-based search-keys to rapidly locate given text sequences and their associated waveforms. Logical constraints on the search, such as AND and NOT, are also enabled. A more detailed search is facilitated by providing corpus specific facilities for displaying and reforming certain subsets of the


JFA_JFB_J03
JFB_JFA_J03

Conversation List

start	end	speaker	PLAY		
0:19	0:21	JFA	stereo	mono	十二月二__十日木曜日
0:21	0:22	JFA	stereo	mono	三__時四__十__五__分
0:22	0:25	JFA	stereo	mono	被験者_ツカモト_収録_を_開始_し_ます
0:25	0:26	JFB	stereo	mono	十二月二__十日
0:27	0:28	JFB	stereo	mono	三__時四__十
0:29	0:29	JFB	stereo	mono	五__分
0:30	0:31	JFB	stereo	mono	被験者アオヤマユカコ
0:31	0:33	JFB	stereo	mono	収録_を_開始_し_ます
0:46	0:46	JFA	stereo	mono	つ
0:56	0:56	JFA	stereo	mono	ないすせ
1:12	1:14	JFB	stereo	mono	ズー
1:31	1:32	JFB	stereo	mono	もしもし
1:32	1:33	JFA	stereo	mono	もしも(し_ー)
1:32	1:34	JFB	stereo	mono	あっはい_アオヤマ_です_ー
1:34	1:35	JFA	stereo	mono	ツカモトです_ー
1:35	1:36	JFB	stereo	mono	こんにち(は_ー)ー
1:35	1:37	JFA	stereo	mono	こんにち(は_ー)

Fig. 4. A page displaying more conventional forms of transcription, offering stereo or mono listening per utterance, as well as showing times for each

various corpora. There is an interface whereby specific combinations of speaker and text type can be entered as search keys and the search constrained by e.g., interlocutor type, gesture dynamics, or discourse mode, by making use of the higher-level annotations on the data.

A Join-Play interactive-editing feature allows the user to simply append the latest utterance segment (video and audio, or audio alone) to a list of related segments to build up a novel data sequence with the speech files and associated text files zipped in a form ready to burn to DVD for wider distribution. This facility has proved useful in the rapid provision of materials for perceptual experiments as well as providing topic-specific subsets of the corpus for analysis at a separate site.

4 Display of Multi-modal Metadata

An increasing amount of our data is multi-modal. We now use 360-degree cameras as well as regular video when recording fresh dialogue data, and use computer programmes to produce derived data from the aligned video and audio. Figure 5 shows the above-mentioned download facility being used with the output of one of the recordings illustrated in Figure 1,

Figure 6 shows how the social dynamics of a multi-party interaction can be readily visualised using colour bar plots for each utterance. It is clear how ‘green’ and ‘grey’ dominate in the first part of this minute, and ‘red’ and ‘yellow’ take over in the latter part, after all four speakers being active simultaneously in the first few seconds. It can immediately be seen, for example, that the feedback provided by red to yellow’s utterances is very different in style from that provided by grey to green. The areas around ‘transition points’ (such as the one

Fig. 5. Selected portions of a conversation can be downloaded for concatenation into fresh dialogues. A time window can be selected interactively to include portions before and after the transcribed segment. A karaoke style subtitle provides visual display of the speech

marked by the cursor in the top part of the figure at time 28.5) can also be of great interest. These are immediately recognisable when scrolling through such graphical displays of speech interaction.

Similar interaction plots, automatically derived, can be related to the video sequence using Flash software for dynamic effect. Figure 7 shows how similar colour-coding can help to identify the movements of each speaker when numerical data are plotted. Here, in offline preprocessing, face-detection software is employed to find the faces in each image, and then image-flow technology tracks the movements around and below each face to provide a track of body and head movements for the different speakers. The derived metadata is displayed in the same clickable form as the text.

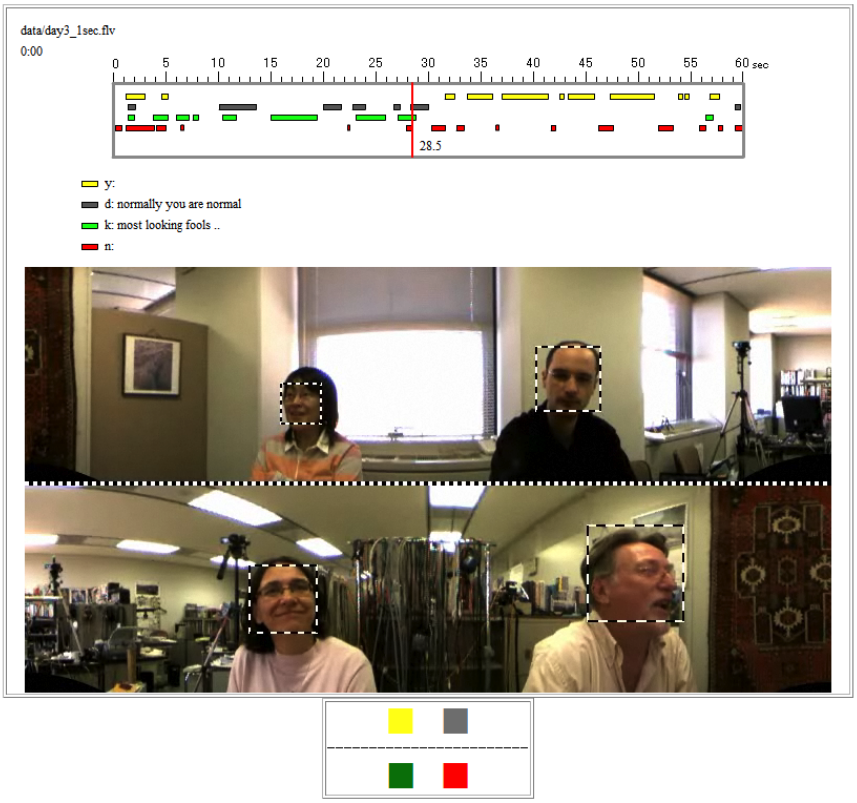

Fig. 6. Another method of interacting with multimodal data. Colour bars show the dynamics of the interaction and a moving cursor marks the point in the discourse, with transcriptions of each talker provided in animated form beneath

5 Conclusion

This chapter has described software for the display of large-corpus data. The web-based tools and interface are now being used by a small community of international researchers working with the dialogue data. However, because of the large amount of personal information included in these highly natural conversations it is not feasible to make the entire corpus publicly available at this time. Representative samples can be seen at the FEAST web pages (Feature Extraction & Analysis for Speech Technology) [16], and interested researchers should apply to the author for access to specific subsets for research purposes. The software, however, can be made freely available to interested researchers with similar data in the hope that standards might then emerge for the interfacing of different types of discourse materials for future technology research and development.

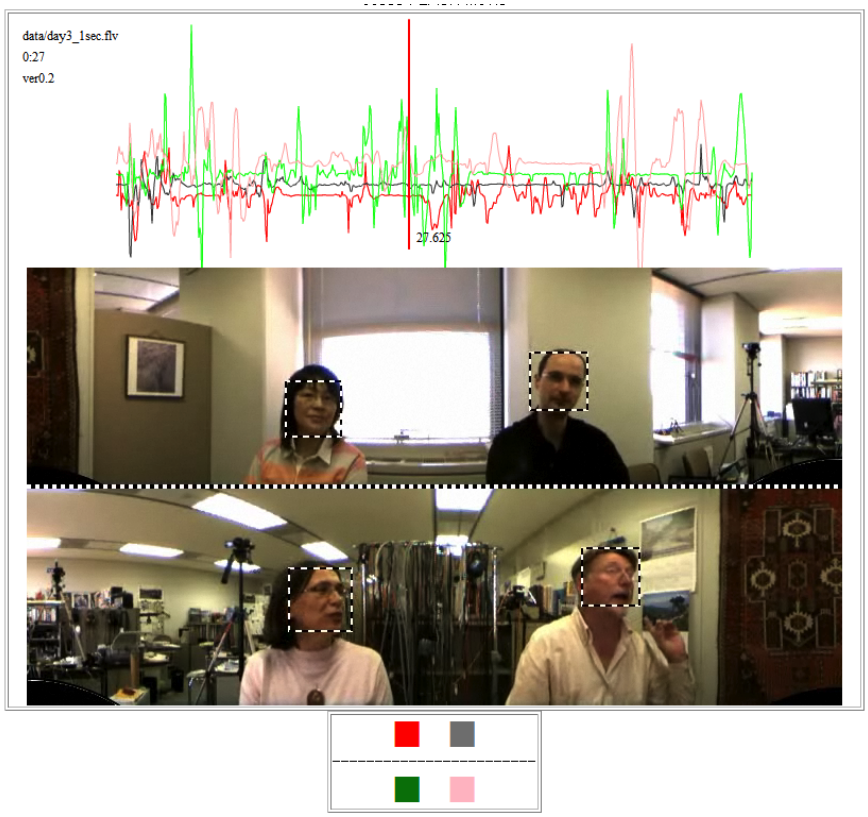

Fig. 7. Using Flash software to animate the automatically-derived movement data. Here the vertical head movement estimates are displayed time-aligned with the video. The cursor can be scrolled through each page of display to control the video and audio output. Again, the dynamics of the interaction become apparent and rapidly guide the researcher.

References

1. ELDA - Evaluations and Language Resources Distribution Agency, Home Page http://www.elda.org
2. LDC - The Linguistic Data Consortium, Home Page http://www.ldc.upenn.edu/
3. Campbell, N., "On the Use of Nonverbal Speech Sounds in Human Communication", pp.117-128,Verbal & Nonverbal Communication Behaviours, Eds A. Esposito et al, LNAI 4775, Springer, 2007
4. ATR: The Advanced Telecommunications Research Institute, Keihanna Science City, Kyoto, Japan.
5. The NIST Rich Transcription Evaluation Project, Meeting Recognition Evaluation, Documentation. http://www.nist.gov/speech/tests/rt/rt2002/
6. Carlette, J., et.al., "The AMI Meeetings Corpus", in proc Symposium on Annotating and Measuring Meeting Behaviour, 2005.

7. Waibel, A., Steusloff, H., and Stiefelhagen, R., "CHIL - Computers in the human interaction loop", 5th international workshop on image analysis for multimedia interactive services, Lisbon, April 2004. Figure 10: Labeller agreement on annotation of changing levels of rapport throughout a conversation
8. Douxchamps, D., Campbell, N., "Robust real-time tracking for the analysis of human behaviour", pp.1- 10 in Machine Learning for Multimodal Interaction, MLMI 2007, LNCS 4892, Springer, 2008.
9. Tucker, S. and Whittaker, S. "Accessing Multimodal Meeting Data: Systems, Problems and Possibilities", in proc Multimodal Interaction and Related Machine Learning Algorithms, Martigny, Switzerland, 2004.
10. Cremers, A. H. M., Groenewegen, P., Kuiper, I., and Post, W., "The Project Browser: Supporting Information Access for a Project team", in proc HCII 2007.
11. Rienks, R.,Nijholt, A., and Reidsma, D., "Meetings and Meeting Support in Ambient Intelligence", in Mobile Communication series, pp.359-378, ch.17, Artech House, ISBN 1-58053-963-7,, 2006.
12. Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turntaking for conversation. Language, 50, 696-735.
13. Local, J., "Phonetic Detail and the Organisation of Talk-in-Interaction", in Proceedings of the XVIth International Congress of Phonetic Sciences. Saarbruecken, Germany: 16th ICPhS, 2007.
14. Campbell, N., "Synthesis Units for Conversational Speech" in Proc Acoustic Society of Japan Autumn Meeting, 2005.
15. SWISH-E – Simple Web Indexing System for Humans, Enhanced Version: http://swish-e.org/
16. FEAST - http://feast.atr.jp/non-verbal/project/html_files/taba/top.html