Tools & Resources for Visualising Conversational-Speech Interaction

Nick Campbell

NiCT/ATR-SLC
National Institute of Information and Comunications Technology
& ATR Spoken Language Communication Research Labs
Keihanna Science City, Kyoto 619-0288, Japan
nick@nict.go.jp, nick@atr.jp

Abstract

This paper describes tools and techniques for accessing large quantities of speech data and for the visualisation of discourse interactions and events at levels above that of linguistic content. We are working with large quantities of dialogue speech including business meetings, friendly discourse, and telephone conversations, and have produced web-based tools for the visualisation of non-verbal and paralinguistic features of the speech data. In essence, they provide higher-level displays so that specific sections of speech, text, or other annotation can be accessed by the researcher and provide an interactive interface to the large amount of data through an Archive Browser.

1. Introduction

With ever-growing increases in the amount of data available for speech technology research, it is now increasingly difficult for any one individual to become 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 are being used in the 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 it's 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 difficult) 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 [10] are very difficult to infer from text alone, and almost impossible to search for without specific and expensive further annotation of the transcription.

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. 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.

2. Browser Technologies

With the growing recent interest in processing multimodal interaction, beginning with projects such as NIST Rich Transcription [3], AMI [4], and CHIL [5], there has been considerable research into collecting and annotating very large corpora of audio and visual information related to human spoken interactions [6], and subsequently huge efforts into mining information in the resulting data [7] and making the information available to researchers from various related disciplines [8]. Consequently, much research has also been devoted to interface and access technologies, particularly using web browsers [9].

Our own corpora illustrate different forms of spoken dialogue and are related by contextual features such as participant identity, 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.

An example is given in Figure 1 which shows the top-level page for one section of the corpus. The page provides access to all the conversations from each participant, grouped in this case according to serial order of the dialogue sequence. Other pages (not illustrated) provide access to the same data grouped according topic of conversation, and by familiarity of the participants, etc.

3. Browsing Dialogue Structure

Whereas complete manual transcriptions are available for most conversations in the corpus, the difficulty of time-aligning such texts is well known to conversation analysts

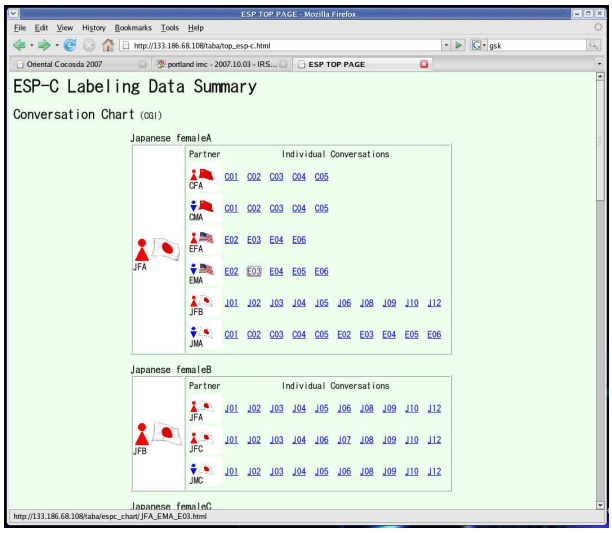

Figure 1: A top page for data access, showing the conversations grouped by speaker and various partners

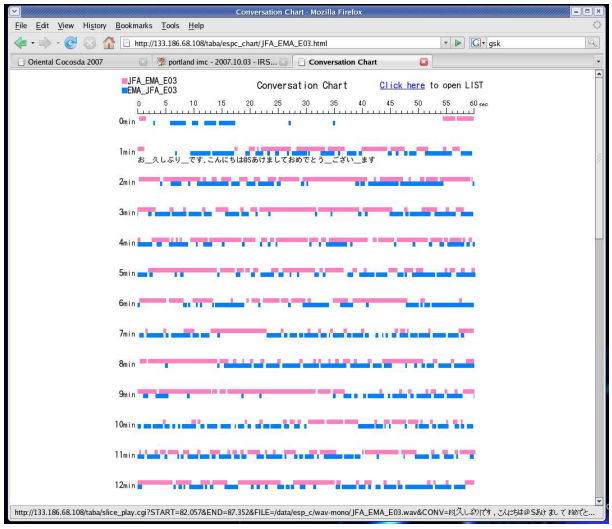

Figure 2: Detail of a sample conversation. This screenshot shows two speakers time-aligned. Mousing on a bar will show the text of that utterance; clicking on the bar will play the speech wave associated with the utterance

(e.g., [11, 12]) who have devised orthographic layout conventions that illustrate (to some extent) the timing and sequential information of the dialogues. We took advantage of the graphical interface of an interactive web page to plot utterance sequences for maximal visual impact as shown in Figure 2. 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 beneath (see e.g., the first row in the figure) and clicking on it plays the speech. This graphical form of layout makes it particularly easy to search utterance sequences based on dialogue structure and speech overlaps.

Two modes of dialogue speech output are offered. Since it is sometimes better to hear a stereo recording allowing access to both speaker's overlapping segments, and other

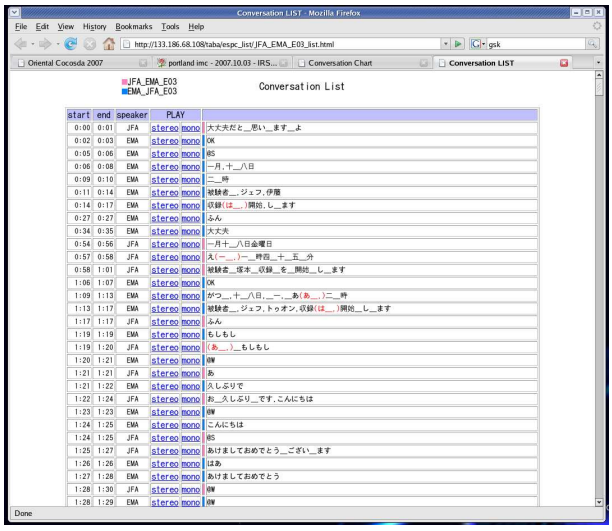

Figure 3: Detail of the aligned transcription allowing direct access to mono or stereo versions of the speech

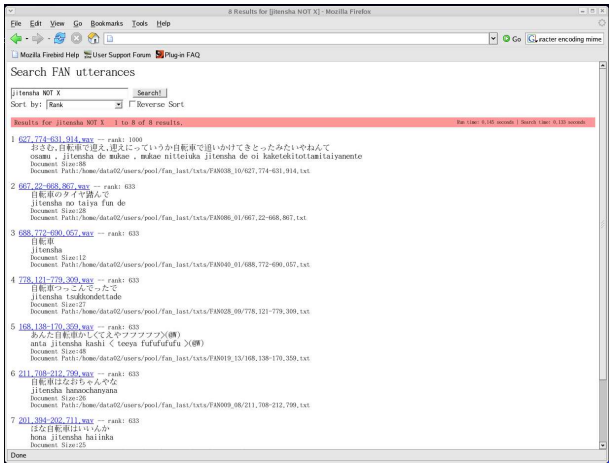

Figure 4: The SWISH-E interface to the corpus

times better to hear a mono version instead, to enable clear listening to an individual utterance in isolation, both forms of speech data replay are made available from a further page (shown in Figure 3) where the whole text of each discourse is displayed in vertical alignment.

Search is an essential facility for any corpus, and several ways are offered for limiting the displayed data to specific subsets. Figure 4 shows a fast Google-type search output, reported in [13] based on the Swish-E public-domain search-engine [14] 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 various corpora. Figure 5 shows 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, or discourse mode, making use of the higher-level annotations on the data.

Novel conversations can be created for use e.g., in perception experiments, and selected samples can be exported to

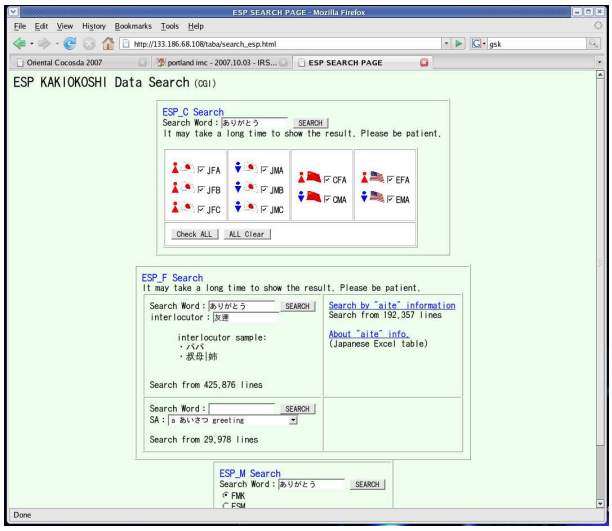

Figure 5: Screenshot of a search window, enabling the user to select a subset of target utterances by combinations of various search parameters

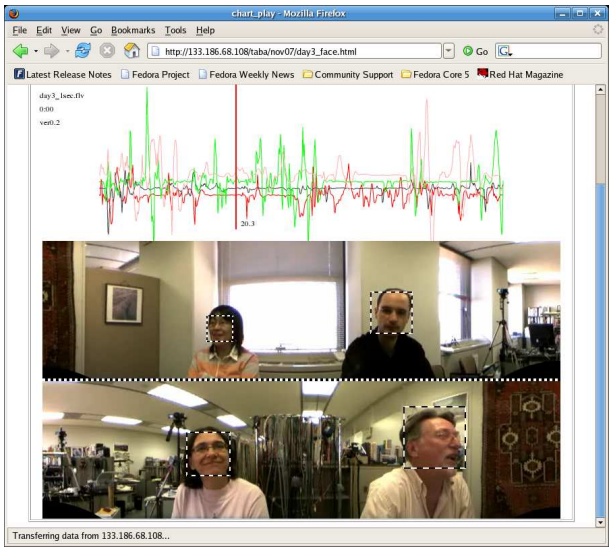

Figure 6: The video playback screen (360-degree lens), with an indicator scrolling through the computer-derived activity tracking for each utterance participant

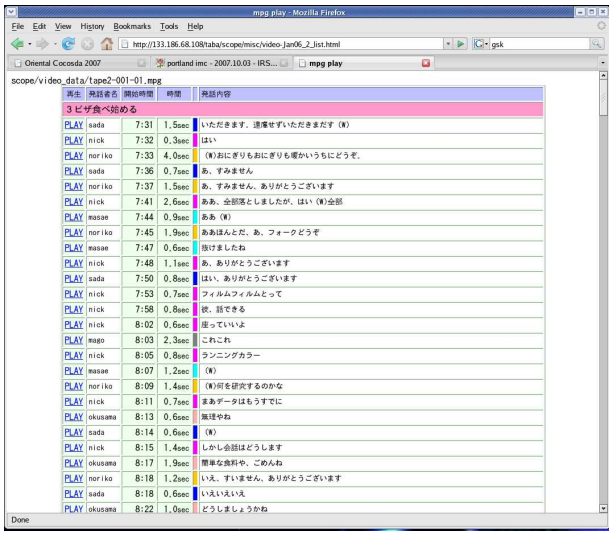

Figure 7: Screenshot showing the aligned transcription of multi-party conversations, with different colours used to identify the different speakers

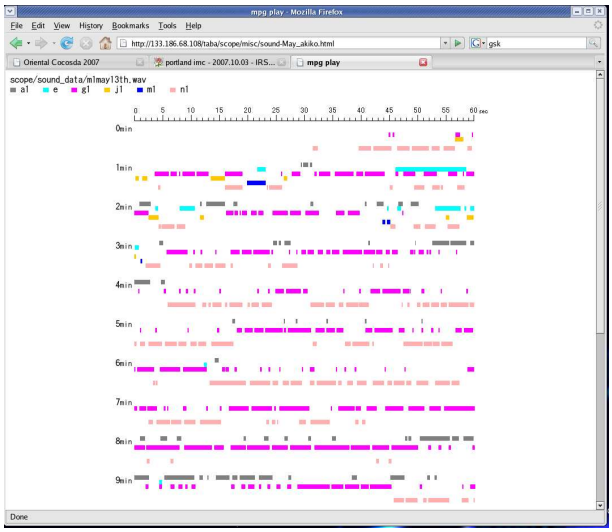

Figure 8: Aligned transcriptions of multi-party conversation showing discourse-level interactions and speaker participation. (Mouse behaviour as for Fig.2)

create a novel sub-corpus with the speech files and associated text files zipped in a form ready to be burned to DVD for wider distribution. 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.

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. Figures 6, 7, and 8 show transcription and plots of such multi-party data. Figure 6 shows how movement plots are related to the video sequence using Flash. Figures 7 and 8 illustrate the use of colour-coding to identify the different speakers. The derived metadata (Figure 9) is displayed in the same clickable form as the text. Figure 10 shows an example of manual annotations of conversational activity (here from 3 labellers) to facilitate e.g., estimates of data reliability.

5. Conclusion

This paper 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. Because of the large amount of personal information included in this highly natural conversational-speech data, it is not possible to make the entire corpus publicly available, but samples can be seen at [15], and interested researchers should apply to the author for access to specific subsets for research pur-

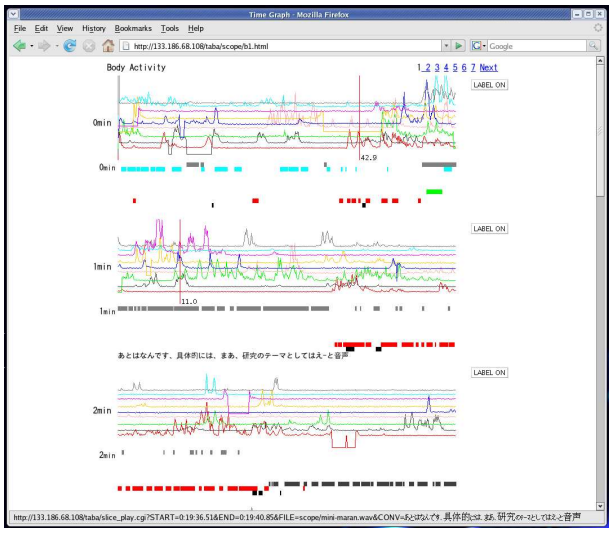

Figure 9: Activity plots for the data shown in Fig7. Here we see the body movements for each speaker aligned by time. Mousing behaves as explained above

Figure 10: Labeller agreement on annotation of changing levels of rapport throughout a conversation

poses. 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..

6. Acknowledgements

This work is supported by the National Institute of Information and Communications Technology (NiCT), and includes contributions from the Japan Science & Technology Corporation (JST), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The author is grateful to Akiko Tabata for her programming, and to the ATR Spoken Language Communication Research Labs for their encouragement and support. An earlier, shorter version of this paper was presented at the Oriental COCOSDA meeting in Hanoi in 2007. The author is also grateful for subsequent discussiona and the reviewers' helpful comments.

References

[1] ELDA - Evaluations and Language resources Distribution Agency, Home Page http://www.elda.org

[2] The Linguistic Data Consortium, Home Page http://www.ldc.upenn.edu/

[3] The NIST Rich Transcription Evaluation Project, Meeting Recognition Evaluation, Documentation. http://www.nist.gov/speech/tests/rt/rt2002/

[4] Carlette, J., et.al., "The AMI Meeetings Corpus", in proc Symposium on Annotating and Measuring Meeting Behaviour, 2005.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50, 696-735.

[12] 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.

[13] Campbell, N., "Synthesis Units for Conversational Speech" in Proc Acoustic Society of Japan Autumn Meeting, 2005.

[14] SWISH-E — Simple Web Indexing System for Humans, Enhanced Version: http://swish-e.org/

[15] http://feast.atr.jp/non-verbal/project/html_files/taba/top.html