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

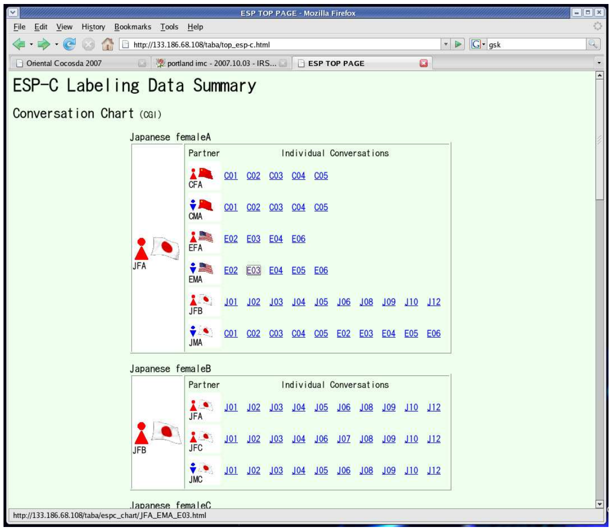
With the growing recent interest in processing multimodal interaction, beginning with projects such as NIST Rich Transcription [3], AMI [4], and CHIL [5], there has been considerable research into collecting and annotating very large corpora of audio and visual information related to human spoken interactions [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


ESP TOP PAGE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/top_esp-c.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
ESP TOP PAGE
http://133.186.68.108/taba/espc_chart/JFA_EMA_E03.html

ESP-C Labeling Data Summary

Conversation Chart (CGI)

Japanese femaleA

JFA	Partner	Individual Conversations
	CFA	C01 C02 C03 C04 C05
	CMA	C01 C02 C03 C04 C05
	EFA	E02 E03 E04 E06
	EMA	E02 E03 E04 E05 E06
	JFB	J01 J02 J03 J04 J05 J06 J08 J09 J10 J12
	JMA	C01 C02 C03 C04 C05 E02 E03 E04 E05 E06

Japanese femaleB

JFB	Partner	Individual Conversations
	JFA	J01 J02 J03 J04 J05 J06 J08 J09 J10 J12
	JFC	J01 J02 J03 J04 J06 J07 J08 J09 J10 J12
	JMC	J01 J02 J03 J04 J05 J06 J08 J09 J10 J12

Japanese femaleC

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

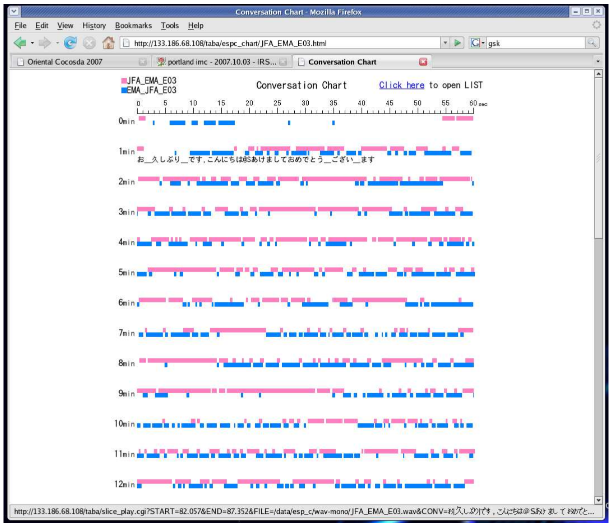
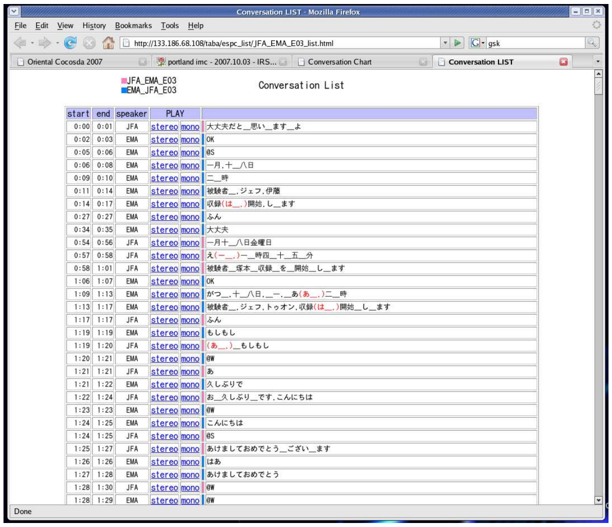

Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/espc_chart/JFA_EMA_E03.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
Conversation Chart
JFA_EMA_E03
EMA_JFA_E03
Conversation Chart
Click here to open LIST
0
5
10
15
20
25
30
35
40
45
50
55
60 sec
0min
1min
お__久しぶり__です,こんにちは@Sあけましておめでとう__ござい__ます
2min
3min
4min
5min
6min
7min
8min
9min
10min
11min
12min
http://133.186.68.108/taba/slice_play.cgi?START=82.057&END=87.352&FILE=/data/esp_c/wav-mono/JFA_EMA_E03.wav&CONV=お|久しぶり|です，こんにちは@Sあけ まし て おめでと...

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

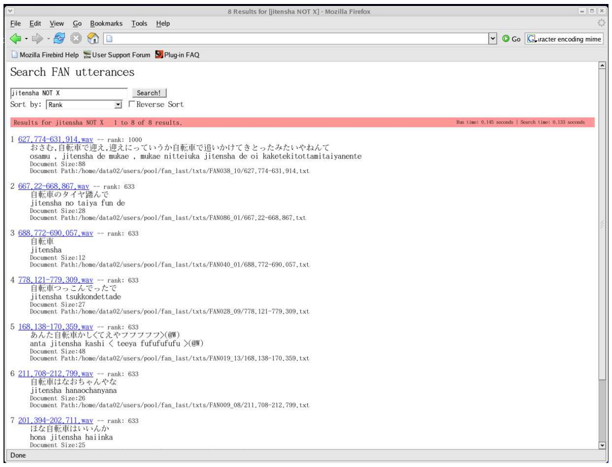

Conversation LIST - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/espc_list/JFA_EMA_E03_list.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
Conversation Chart
Conversation LIST
Done

🟪JFA_EMA_E03
🟦EMA_JFA_E03

Conversation List

start	end	speaker	PLAY		
0:00	0:01	JFA	stereo	mono	大丈夫だと__思い__ます__よ
0:02	0:03	EMA	stereo	mono	OK
0:05	0:06	EMA	stereo	mono	@S
0:06	0:08	EMA	stereo	mono	一月,十__八日
0:09	0:10	EMA	stereo	mono	二__時
0:11	0:14	EMA	stereo	mono	被験者__,ジェフ,伊藤
0:14	0:17	EMA	stereo	mono	収録(は__,)開始,し__ます
0:27	0:27	EMA	stereo	mono	ふん
0:34	0:35	EMA	stereo	mono	大丈夫
0:54	0:56	JFA	stereo	mono	一月十__八日金曜日
0:57	0:58	JFA	stereo	mono	え(一__,)一__時四__十__五__分
0:58	1:01	JFA	stereo	mono	被験者__塚本__収録__を__開始__し__ます
1:06	1:07	EMA	stereo	mono	OK
1:09	1:13	EMA	stereo	mono	がつ__,十__八日,__一,__あ(あ__,)二__時
1:13	1:17	EMA	stereo	mono	被験者__,ジェフ,トゥオン,収録(は__,)開始__し__ます
1:17	1:17	JFA	stereo	mono	ふん
1:19	1:19	EMA	stereo	mono	もしもし
1:19	1:20	JFA	stereo	mono	(あ__,)__もしもし
1:20	1:21	EMA	stereo	mono	@W
1:21	1:21	JFA	stereo	mono	あ
1:21	1:22	EMA	stereo	mono	久しぶりで
1:22	1:24	JFA	stereo	mono	お__久しぶり__です,こんにちは
1:23	1:23	EMA	stereo	mono	@W
1:24	1:25	EMA	stereo	mono	こんにちは
1:24	1:25	JFA	stereo	mono	@S
1:25	1:27	JFA	stereo	mono	あけましておめでとう__ござい__ます
1:26	1:26	EMA	stereo	mono	はあ
1:27	1:28	EMA	stereo	mono	あけましておめでとう
1:28	1:30	JFA	stereo	mono	@W
1:28	1:29	EMA	stereo	mono	@W

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


8 Results for [jitensha NOT X] - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
racter encoding mime
Mozilla Firebird Help
User Support Forum
Plug-in FAQ
Done

Search FAN utterances

jitensha NOT X Search!
Sort by: Rank □ Reverse Sort

Results for jitensha NOT X   1 to 8 of 8 results.   Run time: 0.145 seconds | Search time: 0.133 seconds

```
1 627,774-631,914.wav -- rank: 1000
    おさむ,自転車で迎え,迎えにっていうか自転車で追いかけてきとったみたいやねんて
    osamu , jitensha de mukae , mukae nitteiuka jitensha de oi kaketekitottamitaiyanente
    Document Size:88
    Document Path:/home/data02/users/pool/fan_last/txts/FAN038_10/627.774-631.914.txt

2 667.22-668.867.wav -- rank: 633
    自転車のタイヤ踏んで
    jitensha no taiya fun de
    Document Size:28
    Document Path:/home/data02/users/pool/fan_last/txts/FAN086_01/667.22-668.867.txt

3 688.772-690.057.wav -- rank: 633
    自転車
    jitensha
    Document Size:12
    Document Path:/home/data02/users/pool/fan_last/txts/FAN040_01/688.772-690.057.txt

4 778.121-779.309.wav -- rank: 633
    自転車つっこんでったで
    jitensha tsukkondettade
    Document Size:27
    Document Path:/home/data02/users/pool/fan_last/txts/FAN028_09/778.121-779.309.txt

5 168.138-170.359.wav -- rank: 633
    あんた自転車かしくてえやフフフフフ>(@W)
    anta jitensha kashi < teeya fufufufufu >(@W)
    Document Size:48
    Document Path:/home/data02/users/pool/fan_last/txts/FAN019_13/168.138-170.359.txt

6 211.708-212.799.wav -- rank: 633
    自転車はなおちゃんやな
    jitensha hanaochanyana
    Document Size:26
    Document Path:/home/data02/users/pool/fan_last/txts/FAN009_08/211.708-212.799.txt

7 201.394-202.711.wav -- rank: 633
    ほな自転車はいいんか
    hona jitensha haiinka
    Document Size:25
```

Figure 4: The SWISH-E interface to the corpus

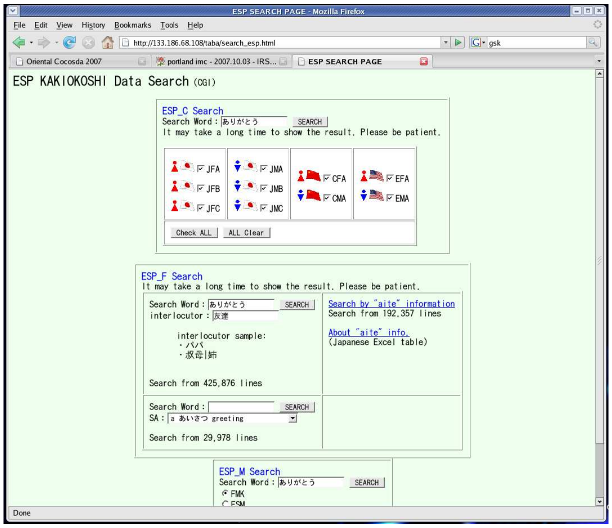
(e.g., [11, 12]) who have devised orthographic layout conventions that illustrate (to some extent) the timing and sequential information of the dialogues. We took advantage of the graphical interface of an interactive web page to plot utterance sequences for maximal visual impact as shown in Figure 2. 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 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


ESP SEARCH PAGE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/search_esp.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
ESP SEARCH PAGE
Done

ESP KAKIOKOSHI Data Search (CGI)

ESP_C Search
Search Word : ありがとう SEARCH
It may take a long time to show the result. Please be patient.

☑ JFA ☑ JMA ☑ CFA ☑ EFA
☑ JFB ☑ JMB ☑ CMA ☑ EMA
☑ JFC ☑ JMC

Check ALL ALL Clear

ESP_F Search
It may take a long time to show the result. Please be patient.

Search Word : ありがとう SEARCH
interlocutor : 友達

interlocutor sample:
・パパ
・叔母|姉

Search from 425,876 lines

Search by "aite" information
Search from 192,357 lines

About "aite" info.
(Japanese Excel table)

Search Word : SEARCH
SA : a あいさつ greeting

Search from 29,978 lines

ESP_M Search
Search Word : ありがとう SEARCH
● FMK
○ FSM

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

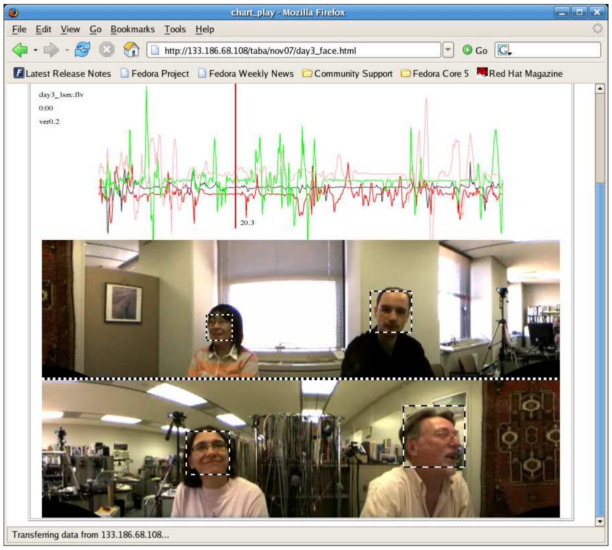
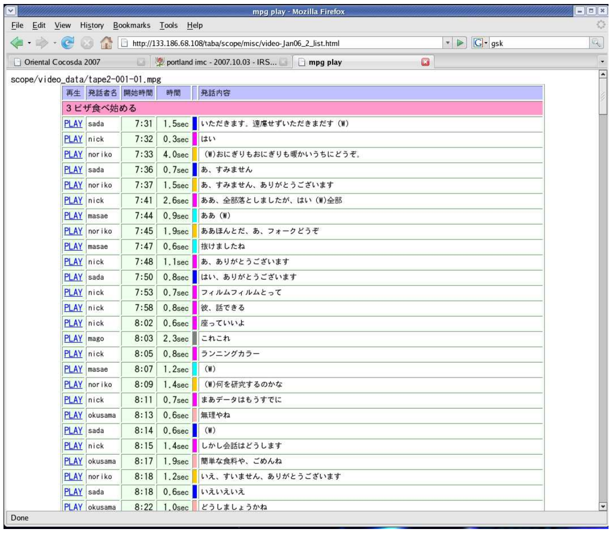

chart_play - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://133.186.68.108/taba/nov07/day3_face.html
Go
Latest Release Notes
Fedora Project
Fedora Weekly News
Community Support
Fedora Core 5
Red Hat Magazine
day3_1sec.flv
0:00
ver0.2
20.3
Transferring data from 133.186.68.108...

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


mpg play - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/scope/misc/video-Jan06_2_list.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
mpg play
Done

scope/video_data/tape2-001-01.mpg

再生	発話者名	開始時間	時間		発話内容
3 ピザ食べ始める					
PLAY	sada	7:31	1.5sec	■	いただきます。遠慮せずいただきまだす (W)
PLAY	nick	7:32	0.3sec	■	はい
PLAY	noriko	7:33	4.0sec	■	(W)おにぎりもおにぎりも暖かいうちにどうぞ。
PLAY	sada	7:36	0.7sec	■	あ、すみません
PLAY	noriko	7:37	1.5sec	■	あ、すみません、ありがとうございます
PLAY	nick	7:41	2.6sec	■	ああ、全部落としましたが、はい (W)全部
PLAY	masae	7:44	0.9sec	■	ああ (W)
PLAY	noriko	7:45	1.9sec	■	ああほんとだ、あ、フォークどうぞ
PLAY	masae	7:47	0.6sec	■	抜けましたね
PLAY	nick	7:48	1.1sec	■	あ、ありがとうございます
PLAY	sada	7:50	0.8sec	■	はい、ありがとうございます
PLAY	nick	7:53	0.7sec	■	フィルムフィルムとって
PLAY	nick	7:58	0.8sec	■	彼、話できる
PLAY	nick	8:02	0.6sec	■	座っていいよ
PLAY	mago	8:03	2.3sec	■	これこれ
PLAY	nick	8:05	0.8sec	■	ランニングカラー
PLAY	masae	8:07	1.2sec	■	(W)
PLAY	noriko	8:09	1.4sec	■	(W)何を研究するのかな
PLAY	nick	8:11	0.7sec	■	まあデータはもうすでに
PLAY	okusama	8:13	0.6sec	■	無理やね
PLAY	sada	8:14	0.6sec	■	(W)
PLAY	nick	8:15	1.4sec	■	しかし会話はどうします
PLAY	okusama	8:17	1.9sec	■	簡単な食料や、ごめんね
PLAY	noriko	8:18	1.2sec	■	いえ、すいません、ありがとうございます
PLAY	sada	8:18	0.6sec	■	いえいえいえ
PLAY	okusama	8:22	1.0sec	■	どうしましょうかね

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

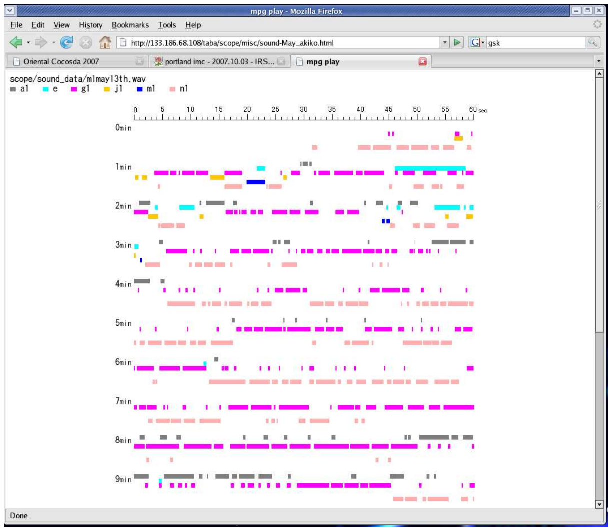

mpg play - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/scope/misc/sound-May_akiko.html
gsk
Oriental Cocosda 2007
portland imc - 2007.10.03 - IRS...
mpg play
scope/sound_data/m1may13th.wav
a1
e
g1
j1
m1
n1
0
5
10
15
20
25
30
35
40
45
50
55
60 sec
0min
1min
2min
3min
4min
5min
6min
7min
8min
9min
Done

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-

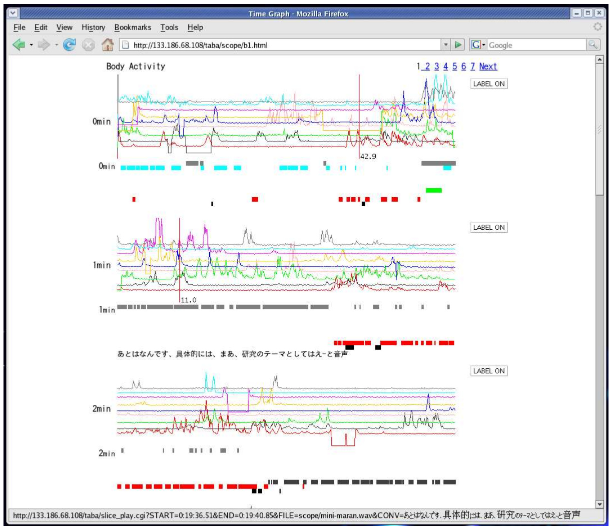

Time Graph - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/scope/b1.html
Google
Body Activity
1 2 3 4 5 6 7 Next
LABEL ON
0min
42.9
0min
LABEL ON
1min
11.0
1min
あとはなんです、具体的には、まあ、研究のテーマとしてはえ-と音声
LABEL ON
2min
2min
http://133.186.68.108/taba/slice_play.cgi?START=0:19:36.51&END=0:19:40.85&FILE=scope/mini-maran.wav&CONV=あとはなんです、具体的には、まあ、研究のテーマとしてはえーと音声

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

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.

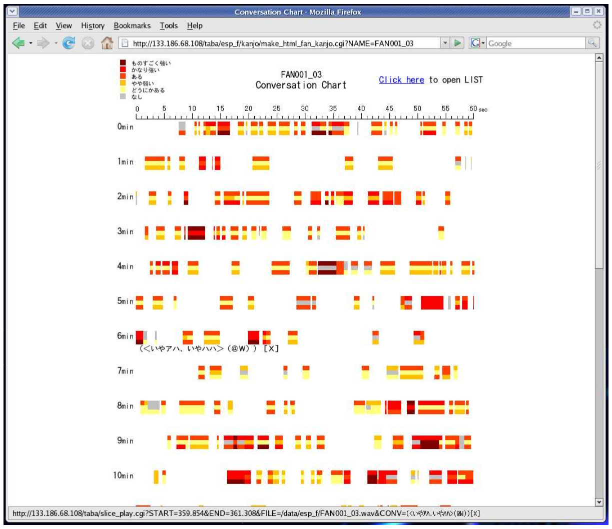

Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://133.186.68.108/taba/esp_f/kanjo/make_html_fan_kanjo.cgi?NAME=FAN001_03
Google
ものすごく強い
かなり強い
ある
やや弱い
どうにかある
なし
FAN001_03
Conversation Chart
Click here to open LIST
0 5 10 15 20 25 30 35 40 45 50 55 60 sec
0min
1min
2min
3min
4min
5min
6min
（<いやアハ、いやハハ>（@W））[X]
7min
8min
9min
10min
http://133.186.68.108/taba/slice_play.cgi?START=359.854&END=361.308&FILE=/data/esp_f/FAN001_03.wav&CONV=(<いやアハ、いやハハ>(@W))[X]

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

[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

Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse

Nick Campbell

National Institute of Information and Communications Technology
& ATR Spoken Language Communication Research Laboratory,
Keihanna Science City, Kyoto 619-0288, Japan.
nick@nict.go.jp

Abstract. This paper describes an analysis of the verbal and nonverbal speaking characteristics of six speakers of Japanese when talking over the telephone to partners of differing degrees of familiarity. The speech data from 100 30-minute conversations between them was transcribed and the acoustic characteristics of each utterance subjected to an analysis of timing characteristics to determine individuality with respect to duration of utterance, degree of overlap, pauses, and other aspects of speech and discourse rhythm. The speakers showed many common traits, but noticeable differences were found to correlate well with degree of familiarity with the interlocutor. Several different styles of interaction can be automatically distinguished in the conversational speech data from their timing patterns.

Keywords: interactive speech, social interaction, nonverbal behaviour, natural data, statistical modelling, real-world applications

1 Introduction

This paper shows that conversational interaction is very much a two-way process, that involves not just the transfer of information related to meaningful content, but also affective displays of attention, involvement, interest, and concern, as well as discourse-level management of the turn-taking and joint processing of the emergent conversational flow. It presents a left-brain/right-brain approach to the processing of conversational speech [1] wherein not just linguistic verbal information but also paralinguistic and nonverbal information is exchanged by means of back-channels and other short non-lexical utterances.

One of the basic and frequently repeated assumptions of conversation analysis is that people talk in turns, and that usually only one person talks at a time [2]. Levinson has defined conversational speech as “a kind of talk in which two or more participants freely alternate in speaking” [3]. This alternation accords well with our sense of the rhythm of a conversation, but it is not well supported by a quantitative analysis of a large number of telephone conversations where people were paid “just to talk” to each other [4]. Goffman [5] differentiates unfocussed interaction, where participants are simply concerned with the “management of sheer and mere copresence”, from focussed interaction where persons “openly cooperate to sustain a single focus of attention". The present study examines data wherein people pass the time by chatting with each other about a variety of topics. Timing details derived from time-aligned transcriptions of these recordings reveal considerable overlapping speech in the discourse. The paper shows that discourse engagement and participation style can be inferred from the structure of these patterns in interactive conversations.

2 A Corpus of Telephone Conversations

The data underlying this study come from a corpus of recorded telephone conversations. One hundred thirty-minute telephone conversations were recorded over a period of several months, with paid volunteers coming to an office building in a large city in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they wore a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48kHz. They did not see their partners or socialise with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis.

In all, ten people took part as speakers in these recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American or Australian English. All conversations were held in Japanese. The non-native speakers were living and working in Japan, competent in Japanese, but not at a level approaching native-speaker fluency. Partners were initially strangers to each other, but became friends over the period of the recordings. There were no constraints on the content of the conversations other than that they should occupy the full thirty-minute time slot. Recordings continued for a maximum of ten sessions between each pair, or five for the non-native speakers.

The speech data were transferred to a computer and transcribed manually using Wavesurfer public-domain speech transcription software [6] to provide a time-aligned record of what was spoken when, by who and to whom. The transcribed speech was aligned at the 'utterance' level to acoustic events in the speech waveform.

The definition of an 'utterance' in conversational speech is problematic. A common practice is to use e.g., any pause in the speech of greater than 200 milliseconds as an objective delimiting boundary, but it was noticed that even many single words contained pauses of more than 300 milliseconds in these conversational data.

It was therefore proposed that our transcribers should use a perception-based "one-yen-per-line" criterion for segmenting the speech, whereby they would increase their payment for more lines produced by cutting the speech into shorter utterances, but would be penalised for breaking up a single utterance into too small or "unnatural" units.

The segmentation was thus largely performed at the level of the phrase, or 'minor intonation unit', i.e., an utterance being a word or group of words demarcated by a single intonation contour. However, in many cases the transcribers actually produced longer and more complex utterance units, with some including punctuation marks such as commas, perhaps because of uncertainty about whether a clearly distinguishing intonational break could be heard or not.

3 Analysis of the patterns of speech activity

A computer program was written to combine and align the separate transcriptions of each speaker's conversations and to calculate the amount of time each person spent silent or talking during the 30-minute sessions.

Four classes of conversational speech activity were thus distinguished from the on/off nature of the speech: (a) both partners silent, (b) both partners talking at the same time, and (c, & d) one or the other partner talking while the other was silent, presumably listening.

Annotations of these speech-activity types were stored in a file along with a record of the length in milliseconds of each utterance, the duration of the pause preceding it, the duration of the previous utterance, and the duration of the pause preceding that. Similar durations were stored for the conversation partner, to facilitate an analysis of the speech activity in general as part of a computer speech processing system for the detection of nonverbal behavior and human participation styles from conversational speech.

The hundred conversations provided 98,698 utterances of between one and fifty syllables in length. 25% of these utterances were less than 500 milliseconds and another 25% longer than 1.5 seconds, with the longest being 11.5 seconds. Median duration of all utterances was 0.9 seconds. Figures 1 to 5 show example sequences. It is clear from the figures that there is often considerable overlap in the conversational speech, and that the definition of a "turn" in such data can be even more problematic than that of an utterance. This point will be addressed further in a subsequent section.

Using the definition of an utterance given above, several utterances can be combined to form a speaker turn. In this implementation, the program counted through each, incrementing if another utterance from the same speaker followed, but resetting the counters whenever the conversation partner started speaking, storing the number of uninterrupted utterances as a parameter in the data table, independently of the duration of any gap between them. A variable indicating whether or not the partner was speaking at the time of onset of the speaker's new utterance was also stored. Table 1 gives details of (a) the number of utterances, and (b) the number of utterances per turn thus derived for the six native-speakers of Japanese in the corpus. By the above criteria, it is clear that by far the majority of turns consist of a single utterance.

The transcribed utterances were then classified into 5 types: (i) 'frequent utterances', i.e., those special speech patterns which appeared more that 25 times each in the transcriptions, (ii) 'Talk', or infrequent content-bearing utterances appearing less than 25 times each, assumed to be more propositional than phatic in content, (iii) laughs, which were subdivided into longer more expressive variants and (iv) shorter more common simple laughs of up to three syllables, and (v) other non-speech noises (grunts) such as sniffs, sharp intake of breath, or coughs which might be used for discoursal purposes. Table 2 shows the distributions of these according to number of utterances in the turn. The table shows a clear difference between distributions for solo speech (in the top part) as against overlapping speech (in the lower part). It also shows a tendency to avoid longer utterances when the partner is talking.

Table 1. Showing the total number of utterances (top) and the number of utterances per turn (bottom) for the six Japanese speakers of the corpus. JFB and JMB spoke only to Japanese partners and so took part in more conversations.

JFA	JFB	JFC	JMA	JMB	JMC
15,543	21,624	13,038	13,122	20,841	11,530

	utts in turn:	1	2	3	4+
JFA	9,386	5,349	2,384	932	721
JFB	12,534	6,724	3,068	1,254	1,488
JFC	8,509	5,394	2,013	694	408
JMA	9,492	6,868	1,807	531	286
JMB	13,408	8,428	3,049	1,117	814
JMC	7,440	4,711	1,718	595	416

Table 2. Showing distributions of utterance types factored according to whether the partner is silent at the onset of speech (top) or still talking (bottom). A clear tendency can be seen for shorter turns (fewer utterances) when the partner is talking. The numbers in the first column show number of utterances in each turn.

	freq	talk	laugh	freq-l	grunt
1	9,988	11,550	1,065	378	285
2	6,773	9,698	900	373	304
3	3,120	3,908	413	179	158
4	1,246	1,514	150	93	81
5	1,063	865	123	77	69

	freq	talk	laugh	freq-l	grunt
1	16,590	15,588	2,298	1082	954
2	1,966	1,731	307	133	119
3	226	184	49	14	14
4	27	25	4	1	1
5	6	4	2	0	0

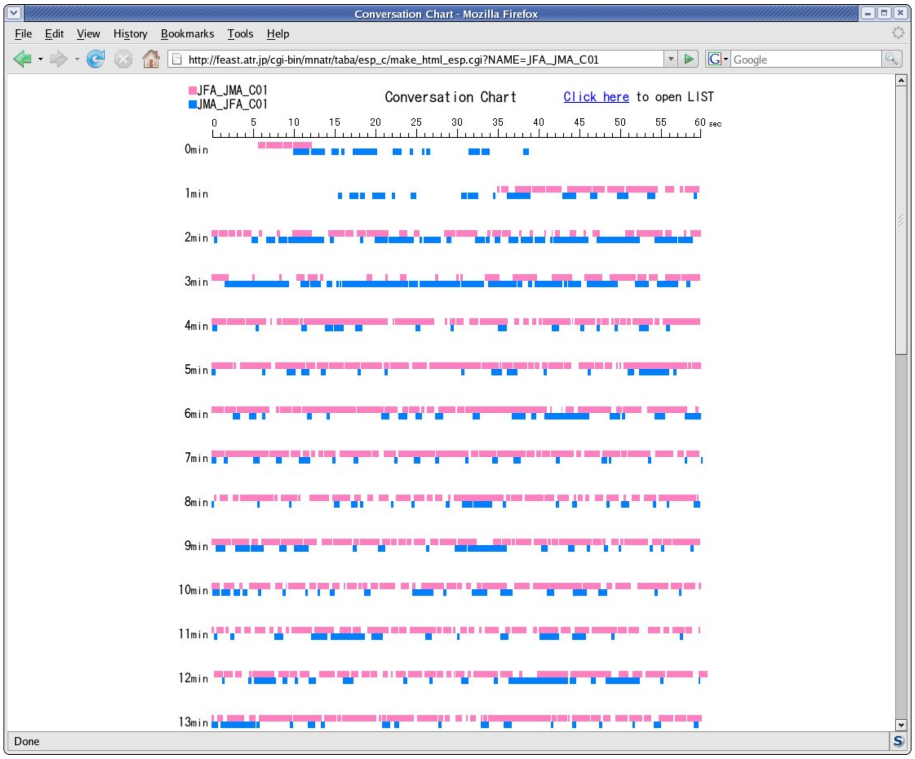

Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/make_html_esp.cgi?NAME=JFA_JMA_C01
Google
JFA_JMA_C01
JMA_JFA_C01
Conversation Chart
Click here to open LIST
0 5 10 15 20 25 30 35 40 45 50 55 60 sec
0min
1min
2min
3min
4min
5min
6min
7min
8min
9min
10min
11min
12min
13min
Done

Fig. 1. A speech activity plot for the first thirteen minutes of the first conversation between Japanese female JFA (pink upper bars in the plot) and her male partner JMA (blue lower bars in the plot). She is considerably older than him and tends to dominate the conversation after the first few minutes. This dominance is clear from the patterns of interaction of speech and non-speech activity. Note how much overlap takes place in their speech from his backchannel activity.

4 Patterns of Speech & Silence

As Figures 1 to 5 show, there is no clear on/off switching between speech turns and silence as might be found if both speakers were using a half-duplex communications channel (as with a 'walkie-talkie' for example), nor is it clear at exactly which point the turn dominance shifts from one speaker to the other. There are clear periods when one partner appears to dominate, but the listener is usually far from passive during these periods. The figures suggest that constant feedback is essential to a conversation and that the nature or style of the conversations can perhaps be characterised by these patterns of speech activity.

4.1 Patterns of Overlapping Speech

In this section we examine in more detail the patterns of overlapping speech of one female speaker from the corpus. Table 3 (from [7]) shows summary durations

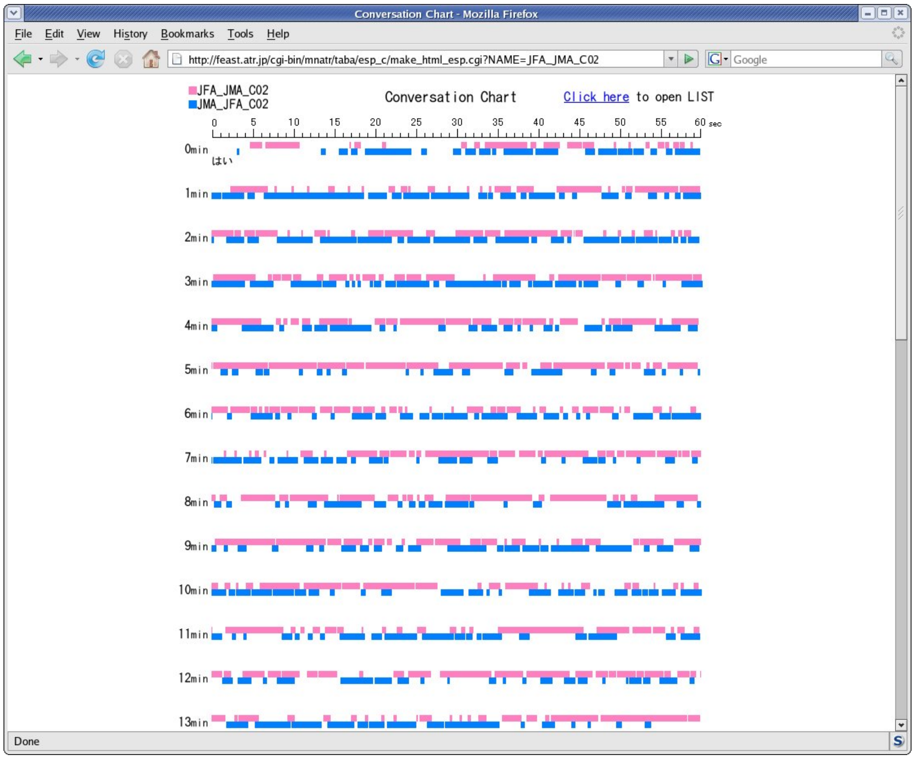


Fig. 2. A speech activity plot for the first thirteen minutes of the second conversation between Japanese female JFA (pink) and her male partner JMA (blue). Note that now the balance is more or less equal between the two partners although there is still considerable overlapping speech. Note too how the relative dominance shifts from one speaker to the other throughout this typical conversation fragment.

in minutes for overlapping speech, solo speech, silence, and total talking time for speaker JFA and her various partners, averaged across thirty conversations. The table differentiates between solo talking, when only one partner is active, overlapping speech, when both are simultaneously active, and silence. Talk time shown is the sum of solo and overlapping speech times. Silence is similarly split between times when one partner is talking (and the other is presumably listening) and those when neither is active. Similar data has been produced for all speakers of the corpus and is summarised in Table 4.

On average she spends 22.7 minutes (sd=2.11) talking during each 30-minute session. Her partners spend on average 17.5 minutes talking with her (sd=2.6). These times sum to more than the total time of each conversation and there is on average 8.6 minutes (sd=2.3) of overlapping speech, with an average of 14.2 minutes of solo speech for JFA with an average of 8.9 minutes of solo speech per partner. Rounding to whole minutes, we find not only that her partners spend the same amount of time in overlapping speech as they do in solo speech but

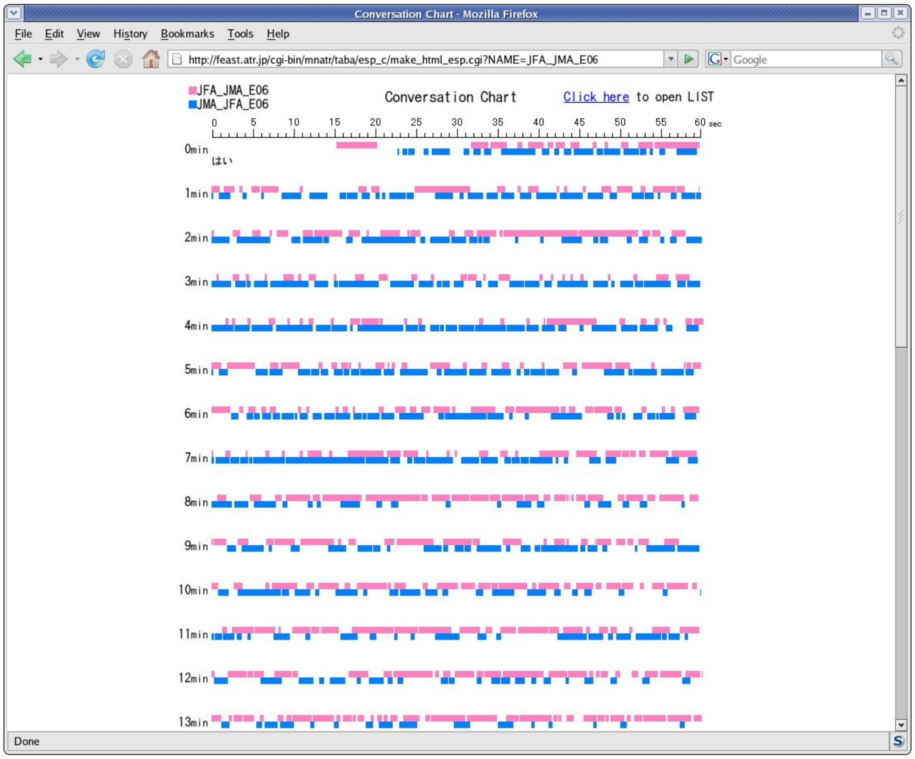

Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/make_html_esp.cgi?NAME=JFA_JMA_E06
Google
JFA_JMA_E06
JMA_JFA_E06
Conversation Chart
Click here to open LIST
0 5 10 15 20 25 30 35 40 45 50 55 60 sec
0min
はい
1min
2min
3min
4min
5min
6min
7min
8min
9min
10min
11min
12min
13min
Done

Fig. 3. A speech activity plot for the first thirteen minutes of the last conversation between Japanese female JFA and male partner JMA. Note that by the final conversation the male clearly dominates throughout the first part of the conversation but there is no evidence of a ping-pong-like exchange of turns.

also that she spends more than 60% of her talking time in overlapped speech. Table 4 (also from [7]) confirms that she is not exceptional. Across the whole range of quartiles for similar data for all speakers, comparing solo talking time to overlapping talking time reveals that all partners spend more than half of their total talking time speaking while the other is also speaking.

4.2 Patterns of Interaction

Taking into account the large amount of speech overlap, this section proposes a measure for categorising the different types of speaker and listener interaction in a dialogue for use in the automatic processing of participant involvement.

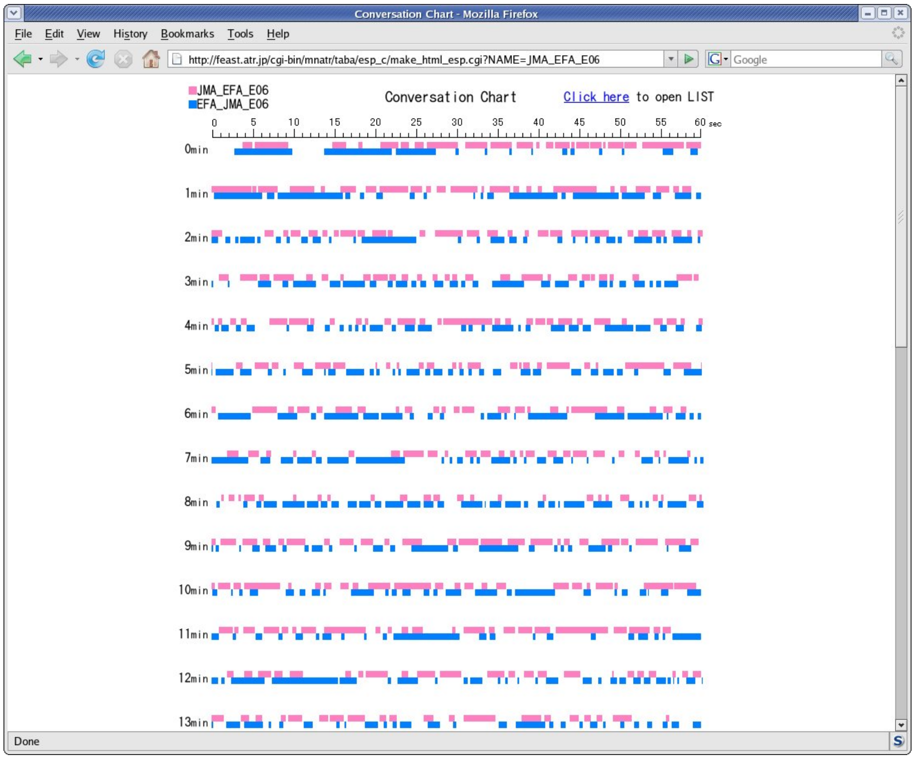
A computer program was written to process the raw transcription files which contain speaker, partner, conversation number, start-time, duration and utterance transcription for each utterance of the corpus. The program calculated for a sliding window of three consecutive utterances the average amount of time spent speaking (from three values) and the average amount of time spent silent

Table 3. Showing mean durations in minutes for overlapping speech, solo speech, silence, and total talking time for speaker JFA (A) and her various partners (B) for their total of 30 30-minute conversations (10 each with the Japanese partners, and 5-each with the non-native-speakers).

	JMA	JFB	CFA	EMA
overlap	8.641	10.932	7.158	5.12
soloA	14.949	12.304	15.6	15.006
soloB	8.247	8.968	8.638	10.308
silA	10.967	10.803	11.06	13.962
silB	17.675	14.138	18.002	18.652
talkA	23.59	23.236	22.758	20.13
talkB	16.888	19.901	15.796	15.428
silent	2.728	1.84	2.422	3.66

Table 4. Showing quantiles summarising speech activity durations for all one-hundred conversations in the corpus. Silence is recorded when neither partner is speaking, overlap when both are speaking at the same time. 'Sil' shows the total time each speaker individually (A or B) was quiet throughout the conversation, presumably while listening. 'Solo' shows the total duration of non-overlapping speech per speaker (A or B), and 'talk' the total overall speech time including overlaps. 'Total' shows timing statistics for the entire conversation (assumed to be 30 minutes by default). All times are shown in minutes. Data are calculated from the time-aligned transcriptions of 100 30-minute conversations

	min	25%	50%	75%	max
silence	0.99	2.08	2.85	3.81	7.03
silA	6.73	10.68	14.02	16.91	22.46
silB	5.72	13.09	14.68	17.68	21.58
soloA	4.14	9.51	11.66	14.68	18.17
soloB	4.55	8.39	10.64	13.32	18.90
overlap	2.66	5.53	7.01	9.04	12.80
talkA	10.80	16.04	18.75	22.44	28.52
talkB	12.20	15.66	17.93	20.15	27.15
total	28.57	32.00	32.93	33.96	37.98



http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/make_html_esp.cgi?NAME=JMA_EFA_E06

Fig. 4. A speech activity plot for the first thirteen minutes of the last conversation between Japanese male JMA (pink, upper bars) and female English-native-speaker partner EFA (blue, lower bars). Note how evenly-balanced their interactions have become by the final conversation. Neither dominates, and their conversation appears much more fragmented than the previous examples, perhaps because of her limited command of the Japanese language.

(from four values) and produced a ratio of the speech time divided by the time not speaking, that was scaled by the duration of the centre utterance, as shown in Equation 1.

$$flow = sd_t * (sd_{t-1} + sd_t + sd_{t+1}/3)/(nsd_{t-1} + nsd_t + nsd_{t+1} + nsd_{t+2}/4) \quad (1)$$

where sd_t represents the duration of utterance at time t, and nsd_t represents the pause duration preceding the utterance at time t.

This measure is high when the speech-to-silence ratio is high, and low when the pauses surrounding an utterance tend to be long. It thus provided a measure of local 'density' of speech activity, or 'flow' of the dialogue. High values occur at times of high information content, and low values at times of backchannel or 'active-listening' activity.

Figure 4 plots histograms of the averaged speech (sp) and silence (gp) durations in the log domain. Original times are in seconds, so for example a value of

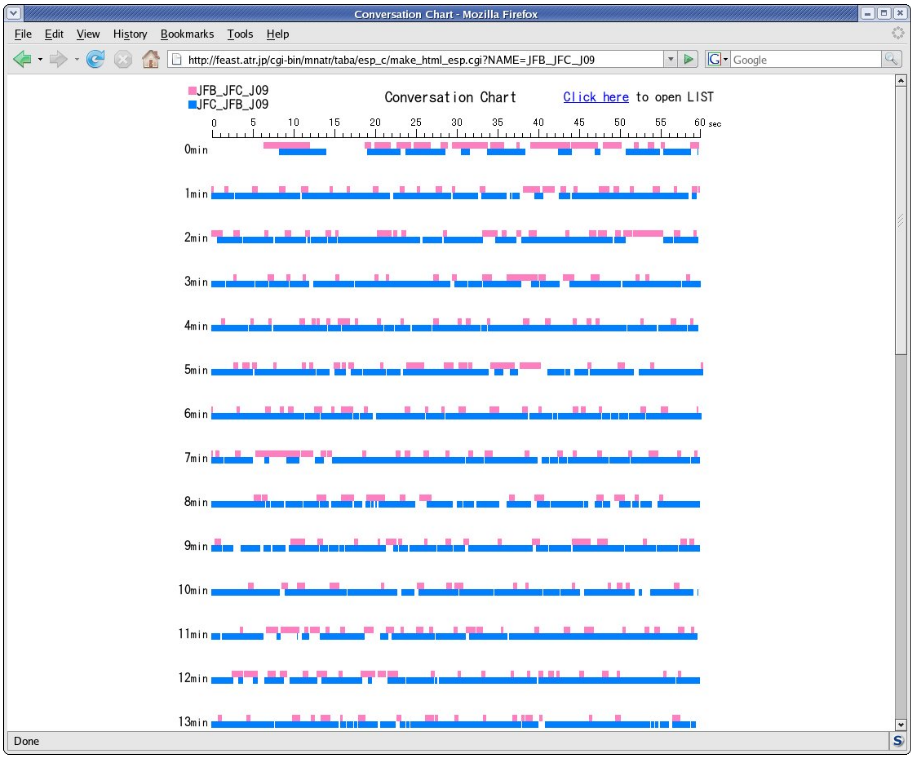

Conversation Chart - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://feast.atr.jp/cgi-bin/mnatr/taba/esp_c/make_html_esp.cgi?NAME=JFB_JFC_J09
Google
JFB_JFC_J09
JFC_JFB_J09
Conversation Chart
Click here to open LIST
0 5 10 15 20 25 30 35 40 45 50 55 60 sec
0min
1min
2min
3min
4min
5min
6min
7min
8min
9min
10min
11min
12min
13min
Done

Fig. 5. A speech activity plot for the first thirteen minutes of the last conversation between Japanese female JFB (pink, upper bars) and Japanese female partner JFC (blue, lower bars). Note how JFC totally dominates the conversation, and how often JFB briefly joins in with her own short contributions.

0 represents a second, 1 represents 2.718 seconds, and -2 135 milliseconds. The values, being durations, plot close to normal on a log scale, but the difference between the upper plot (speech) and the lower plot (silence) shows the tendency for fewer long pauses with a clear skew to the left for the pause durations..

Figure 4 shows boxplots of the average *flow* values for the ten conversations of speaker JFB talking to JFA. A clear trend can be seen for increasing values between the first and fourth conversation, then a reset (for some reason perhaps explained by her catching a cold during the winter break) before another clear and progressive increase from the fifth to the tenth conversation. These trends can be interpreted as representing a shift from more passive interaction in the early stages, to a more active role in the conversations with the progression of time and the increase in familiarity between the two conversants.

Figure 4 plots similar boxplots showing the mutual interactions between JMC talking with JFB (top left), JMC talking with JMB (top right), JFB talking with JMC (bottom left), and JMB talking with JMC (bottom right). Reciprocity

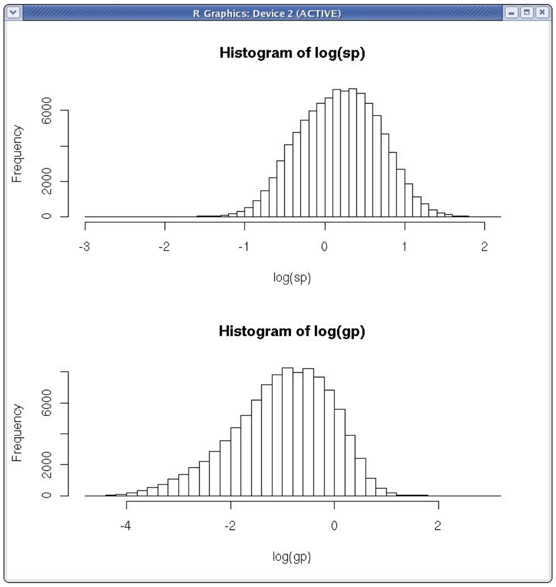


Fig. 6. Histograms of log speech (upper part, sp) and pause (lower part, gp) durations, showing a distribution close to normal for the speech segments, but a distinct skew indicating more short and fewer long pauses

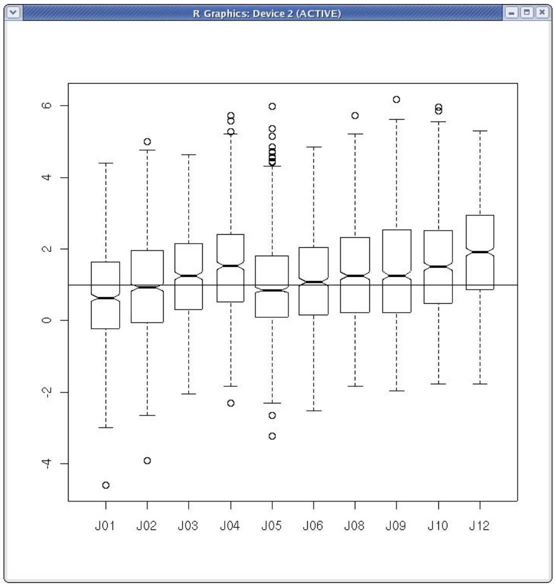


Fig. 7. A boxplot of the flow values for speaker JFB talking with JFA across a series of ten conversations. We can see a clear increase of dominance across the first four conversations, then a reset, then a continued increase with time as they become more familiar.

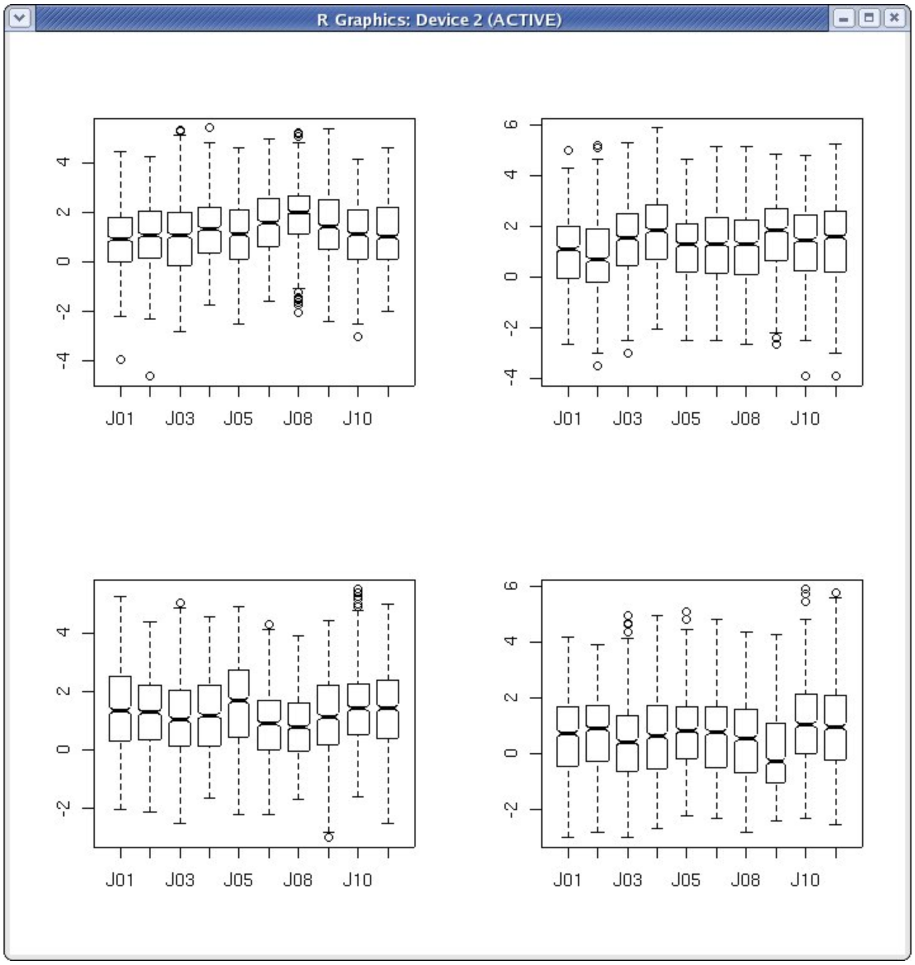


Fig. 8. 4 boxplots showing the mutual interactions between JMC talking with JFB (top left), JMC talking with JMB (top right), JFB talking with JMC (bottom left), and JMB talking with JMC (bottom right). Reciprocity can be seen in both upper-lower pairs as one partner relatively dominates each conversation to a different extent.

can be seen in both upper-lower pairs as one partner relatively dominates each conversation to a different extent.

Table 5 shows the correlations measured between averaged flow measures for all conversations between each pair of speakers in the Japanese native-speaker dialogues. If the 'ping-pong' model of conversational turn-taking is correct, with one partner remaining quiet while the other is speaking, then we would expect all pairs to show similar correlations to that found between male JFB and female JFC. Their high negative correlation of -0.75 indicates that he tends to listen quietly when she speaks, and vice versa. However, of the five remaining pairs of speakers, the table shows only two weak correlations of $r = -0.3$ and three very weak correlations of less than $r = -0.1$, being in fact weakly *positive* for the two male speakers JMC JMB. This indicates that these two men tended to speak when the other was speaking, and to be quiet when the other was quiet. They bonded well. Much the same can be said for all other pairs except JFB and JMC.

Table 5. Correlations between measures of discourse flow for each pair of conversants. The measures show surprisingly little reciprocity.

speakers	correlation
JFB JMC	r= -0.749
JFC JFB	r= -0.314
JMA JMB	r= -0.306
JFB JFA	r= -0.070
JFC JMB	r= -0.010
JMC JMB	r= 0.068

5 Discussion

It appears from this data that a naturally interactive dialogue is not like a tennis match, where there is only one ball that can only be in one half of the court at any given time. Rather it is like a volley of balls being thrown in several directions at once. The speaker does not usually wait silently while the listener parses and reacts to an utterance; there is a constant exchange of speech and gesture, resulting in a gradual process of mutual understanding wherein a consensual 'meeting of the minds' can take place [8].

There is not room in the present paper to discuss the local variations of the flow measure throughout a given dialogue, but this analysis is part of present and future work to automatically determine the level of rapport between participants in a dialogue and to measure on an utterance-by-utterance basis the degree to which they can be said to be involved in the conversation.

6 Conclusion

This paper has presented some results of an analysis of a large body of conversational speech recordings. It has shown that contrary to naive assumptions of dialogue as a tennis-like exchange of question and answer or topic and comment, it actually presents a complex pattern of simultaneous talking as partners take turns to dominate in the interaction. There appear to be no clear boundaries between one turn and the next, and the shift from backchannel feedback to conversational dominance appears to be more subtle.

A measure was proposed that quantifies the degree of participant interaction in a dialogue by estimating the ratio of speech to non-speech and length of utterance versus length of surrounding pauses. This measure can be used to chracterise a conversation in terms of joint activity of the partners, and it was shown that in the majority of partner pairs, both tended to speak simultaneously in many cases.

Acknowledgement

This work is partly 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 especially grateful to the management of Spoken Language Communication Research Labs at ATR for their continuing encouragement and support.

References

1. Springer, S. P. & Deutch, G. “Left brain, right brain: perspectives from cognitive neuroscience”. Fifth edition, W.H. Freeman, N.Y., 1998.
2. Sacks, H., Schegloff, E., A., and Jefferson, G.,. “A simplest systematics for the organization of turn taking for conversation”. In Schenkein 1978
3. Levinson, S., C., *Pragmatics*, Cambridge University Press, 1983.
4. Campbell, N., “Databases of Expressive Speech”, *Journal of Chinese Language and Computing, Vol 14, N.4*, pp 295-304, 2004.
5. Goffman, E., *Behaviour in public places: Notes on the social organisation of gatherings*, Free Press of Glencoe, New York, 1963.
6. Wavesurfer, Open Source tool for sound visualization and manipulation: www.speech.kth.se/wavesurfer/
7. Campbell, N., “Approaches to Conversational Speech Rhythm: Speech Activity in Two-Person Telephone Dialogues”, pp.343-348 in Proc XVIth International Congress of the Phonetic Sciences, Saarbrucken, Germany, 2007.
8. McNeill, D., Quek, F., McCullough, K-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X-F., & Ansari, R. Catchments, Prosody, and Discourse. *Gesture* 1, 9-33. 2001.