'Active Listening' Characteristics Confirm Spoken Dialogue as a Very Two-way Interactive Process *

◎ Nick Campbell, NiCT/ATR
nick@nict.go.jp

1 Introduction

Many interactive speech processing systems consider spoken dialogue to be like a game of tennis, or ping-pong, with a single 'ball' that is thrown back and forth between the partners. In this simple turn-taking model, when the user makes a statement or asks a question, the system responds and then waits for the next utterance from the user. However, real-life spoken dialogue employs no such constraints on turn-taking, and both partners typically interrupt each other frequently in the mutual construction of shared understanding through speech activity.

This paper examines the patterns of interactive speech activity in some conversational speech data collected as part of the JST/CREST ESP project between 1999 and 2005 [1]. It shows that contrary to the above expectations, overlapping speech accounts for as much as half of the individual speaking times in a set of 100 30-minute telephone conversations between paid volunteer participants.

2 Dialogue Fragmentation & Flow

Figure 1 shows plots of speech activity for the first 13 minutes of the last conversation between one particular pair of subjects from the above corpus. There were no constraints on the content of the conversations other than that the partners should talk to each other over a telephone line (with no face-to-face contact) for a period of 30 minutes each week over a period of 3 months.

The figure clearly shows the utterances to be fragmented. In the subsequent transcriptions of the conversations, there were very few complete grammatical sentences, and much more interactive turn-taking with the 'listener' often completing utterances of the 'speaker', providing backchannel feedback, and expressing parallel opinion or jointly rephrasing the speaker's current utterance.

Table 1 gives summary statistics for the entire set of conversations between all pairs of Japanese native-speaker subjects. Data are calculated from the time-aligned transcriptions of 100 30-minute conversations. Silence is noted when neither partner is speaking, overlap when both are speaking at the same time. All times are shown in minutes. The table shows median silence duration to be a little over 14 minutes for each partner, and median speaking time, including overlaps, to be around 18 minutes for each partner. Talking time adds up to more than the median conversation time of 33 minutes. Overlapping speech occurs for 7 minutes (median). This is much more than half of the median solo talking time for each partner.

Table 1 Showing quantiles summarising speech activity durations for all one-hundred conversations in the corpus. '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).

	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

3 A Measure of Discourse Flow

An effective measure of this type of 'discourse flow' can be obtained using the following formula,

$$flow = sd_t * \frac{0.333 * \sum_{i=t-1}^{t+1} sp_i}{0.25 * \sum_{i=t-1}^{t+2} nsp_i}$$

where sd_t is the duration of the current speech fragment, sp_i is the smoothed duration of current and neighbouring speech fragments and nsp_i is the smoothed duration of neighbouring non-speech periods.

This produces a ratio of speech to non-speech activity scaled by the length of the current utterance. When high, it indicates that the speaker is dominating the discourse, and when low it may be taken as a sign that the speaker is spending more time listening or thinking than speaking.

If conversation follows a ping-pong pattern, as if using a push-to-talk switch or an asynchronous line, there will be a very high negative correlation between the flow measures for each speaker across a set of conversations. A positive correlation would occur if both partners tended to speak (or remain silent) at the same time.

Table 2 shows the correlation of flow measures between ten sets of 30-minute conversations between six pairs of speakers in the ESP_C corpus. It reveals

---

*アクティブ リスニング、音声対話の双方向的特徴について
ニック キャンベル NiCT/ATR-SLC, Keihanna Science City, Kyoto, 619-0288, Japan,

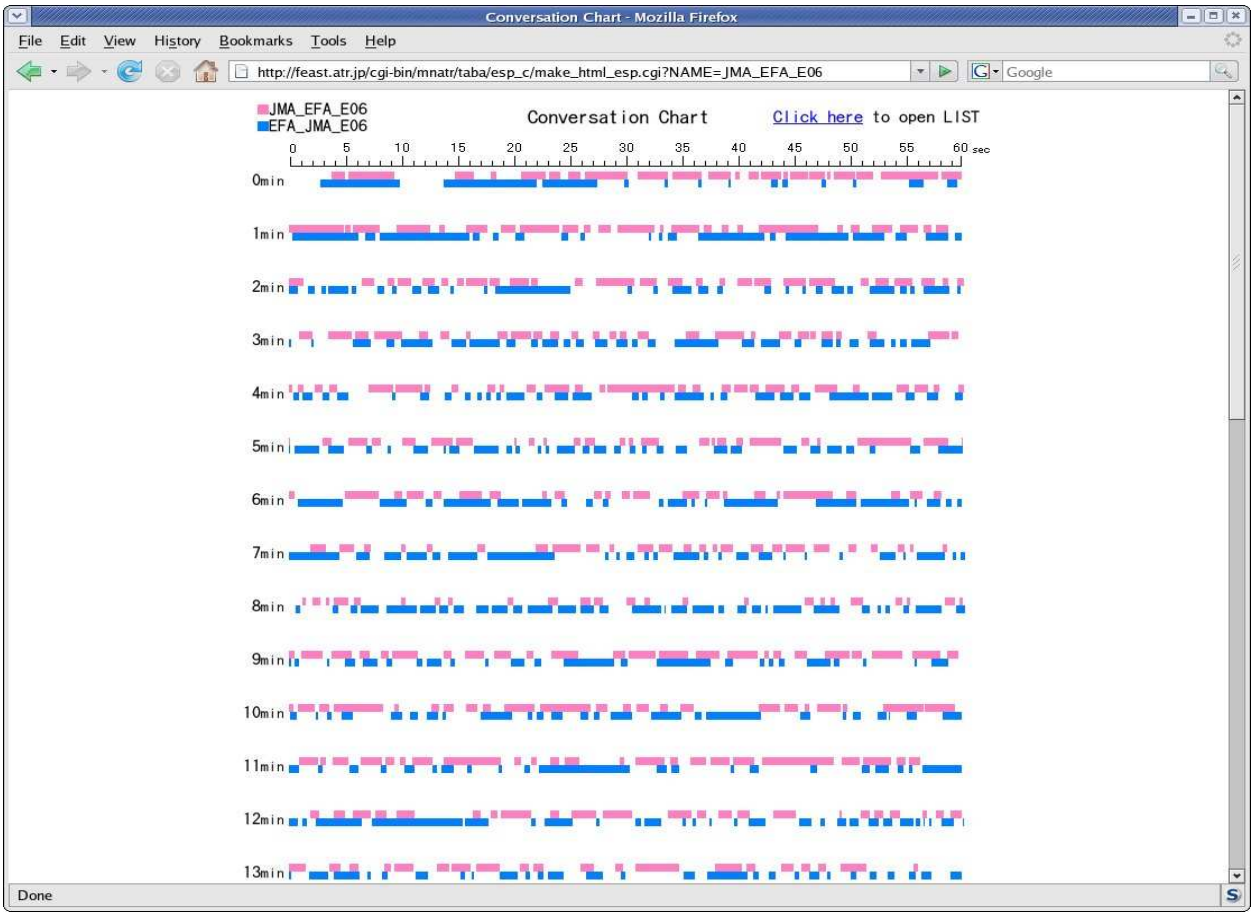

Fig. 1 A speech activity plot for the first thirteen minutes of the last conversation between Japanese male JMA (upper bars) and female English-native-speaker partner EFA (lower bars). It is clear who is the dominant speaker at any given moment, but the partner is also very active, with frequent feedback and additive contributions

that only partners JFB & JMC preferred to alternate their utterances, with one being quiet while the other spoke, resulting in a high negative correlation. Partners JFC & JFB (both female) and JMA & JMB (both male) showed a much smaller negative correlation, indicating more joint activity. But JFC & JMB, and JMC & JMB had no negative correlation in measures of flow across their dialogues, with the latter pair having a small *positive* result. JMB, a young male, appears to be quite garrulous and enthusiastic, especially with JMC with whom he shared some very lively conversations.

Taken together, these results indicate that short, fragmented utterances are common in spoken conversations and it can be inferred from this that, as with the 'handshaking' of modems, frequent feedback may be necessary for a more efficient flow of communication.

Table 2 Correlations between measures of discourse flow for each pair of conversants. The measures show surprisingly little reciprocity.

r =	JFB & JMC -0.749	JFC & JFB -0.314	JMA & JMB -0.306
r =	JFB & JFA -0.070	JFC & JMB -0.010	JMC & JMB +0.068

4 Conclusion

This paper has presented some data from the ESP_C corpus of expressive telephone conversations and shown that contrary to the expectations of polite discourse, there is considerable overlap of speech segments as both participants in the conversation collaboratively build a shared understanding. The 'listener' is often as active as the 'speaker' in these conversations, providing feedback and elaboration of the dialogue.

If such discourse styles are to be modelled in future spoken-dialogue interfaces, e.g., with robots or information-provision systems, then the assumption of a push-to-talk type of turn-taking will not be sufficient. For 'Active-Listening' technology, a speech synthesiser would require a means of sensing the participation status of any listener(s) present, and means of adjusting its speech output to maximise the communication efficiency. This work is currently in progress, using technology developed as part of the SCOPE "Robot's Ears" project [2].

References

[1] JST/CREST Expressive Speech Processing project, introductory web pages at: **http://feast.atr.jp/esp/**

[2] Processing Participation Status in Meetings, SCOPE-sponsored research: **http://feast.atr.jp/nonverbal}**