interspeech 2008 incorporating SST 08

Exploring a Mechanism of Speech Sychronization Using Auditory Delayed Experiments

Masato Ishizaki[1], *Yasuharu Den*[2] *and Senshi Fukashiro*[3]

[1]Graduate School of Interdisciplinary Information Studies, The University of Tokyo
[2]Faculty of Letters, Chiba University
[3]Graduate School of Arts and Sciences, The University of Tokyo
ishizaki@iii.u-tokyo.ac.jp, den@cogsci.l.chiba-u.ac.jp, fukashiro@idaten.c.u-tokyo.ac.jp

Abstract

This paper investigated how speakers synchronize their speech by experiments in which the participants naturally and simultaneously recited under auditory delayed conditions. Statistical analysis revealed that the speakers changed strategies to adjust the timing of their utterances. This finding constitutes one fundamental mechanism for coordinating verbal behavior that can contribute to designing comfortable interactions with virtual agents or robots.

Index Terms: joint actions, speech synchronization, auditory delayed experiments

10.21437/Interspeech.2008-15

1. Introduction

As speech, language, image, and sensor technologies have advanced, a variety of human-computer or human-robot interaction systems have been developed. The research foci have been diversified from speech to multi-modal, from two-party to multi-party, and from content to trust [3, 11]. In this context, the research findings in other fields such as cognitive and social psychology are being reconsidered to improve the interface with virtual agents or robots.

Clark (1996) proposed that conversation is a joint activity based on speech act theory [2, 15]. He provided supporting evidence by examining his data from the speech to understanding levels. For example, speakers distinguish the expressions "um" and "uh" depending on their expectation of the length for subsequent silences [7]; monitoring the interlocutor's activities and exchanging information on a moment-by-moment basis has been shown to be a key for efficient task achievement [8].

One of the most fundamental phenomena in joint activities is the synchronization or the convergence of interactants' behavior. Brennan and Clark (1996) demonstrated that speakers collaboratively conceptualize objects in conversations and share converged expressions for them. Synchronization is also observed in syntactic structures [5], speech rate [4], speech rhythm [9], and bodily movements (postural sway) [16].

Den et al. (2007) modified Shockley's framework [16] to examine the relationship between speech and postural sway in conversations. Their speakers listened to their interactant's delayed speech and studied whether asynchronous speech affected the synchronization of the postural sway. They showed that synchronized behavior is statistically smaller in the delayed speech condition than in the normal speech condition without investigating how speakers synchronized their speech and bodily movements.

This paper focuses on how speakers synchronize their speech by examining the patterns of adjusting the timing of utterances based on the auditory delayed experiments in Den et al. (2007). The paper is organized as follows. The next section summarizes the experimental settings, Section 3 describes the analysis of the experimental results, and Section 4 discusses the implication of the findings.

2. Experiments

2.1. Participants

Twenty-two pairs (9 males and 13 females) of undergraduate and graduate students in Tokyo participated in the experiments. All pairs were of the same gender.

2.2. Task

The participants were told to jointly find a possible train route from the departure station to the arrival station on their maps by exchanging information. There is only one possible route, and their maps are different from each other: There are some stations and/or their connecting routes on one map, but not on the other. Hence, the participants need to provide and obtain each other's information on the stations and the routes. In the end of the conversation, they were requested to simultaneously recite the route they found for verification. As the task here is a variant of the Map Task [1], the participants were spontaneously engaged in the conversations.

Three kinds of fictitious train maps were prepared using ten stations and routes. The station names were chosen from the prefectural capitals in Japan. The station names and routes were printed in color in A2 size and the font size was 24 point. The maps were displayed two meters from the participants at eye level.

2.3. Equipment

The participants talked with each other wearing a headset microphone (Sennheiser HMEC322 + ME104) and heard their own voices without delay from the left channel and the interactant's voice from the right channel with a delay of 0, 250, or 500 msec set by the experimenters. The speech delay was created using equipment that generates an auditory delay (Roland RDL-2040). The speech was recorded on a laptop PC using an audio capture device (Edirol UA-1000).

The participants' bodily movements were recorded using a video camera (Sony VX-2000) and a motion capture device (Vicon MX) at a sampling frequency of 50 Hz. Ten markers were attached to the participants: three each for the head, the chest, and the waist, and one for identification. The signals from the

Accepted after peer review of full paper

September 22 – 26, Brisbane Australia

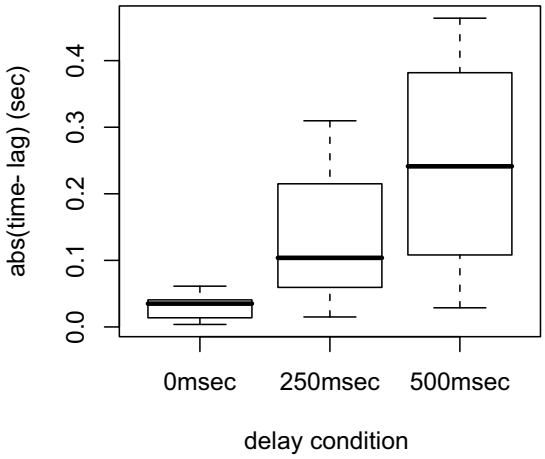

Figure 1: *End time differences between speakers*

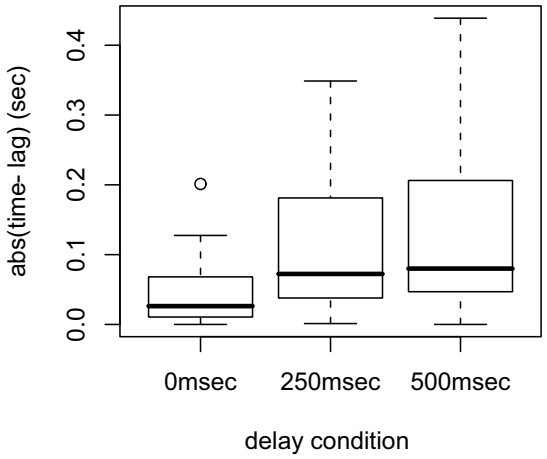

Figure 2: *Start time differences between speakers*

video camera and from the motion capture device were synchronized and recorded on two laptop PCs each for the participants.

2.4. Procedure

1. The participants completed questionnaires, and the markers for the motion capture device were attached.
2. The experimenters explained how to solve the task using the written material. Then the participants solved an example task on trial to make sure they understood.
3. The participants stood back to back inside separate sound insulating panels and wore headset microphones. The panels were set in the center of the room.

Three experiments were carried out for each pair with three different maps and three delay conditions: 0, 250, and 500 msec. The maps and the delay conditions were counterbalanced. The participants were asked to complete the task by simultaneously reciting the stations from the start to the end found in the conversation.

They took a three to five minute break between experiments. After all the experiments were finished, the correct routes and the purpose were explained to the participants.

3. Data Analysis

To examine how the speakers synchronized their speech, simultaneous recitation of the correct route section was analyzed. After excluding pairs who did not recite the route and had difficulty solving the task, 15 of the 22 pairs remained for analysis.

A phonetic transcription of the conversations shown in Table 1 was done by one of the authors for subsequent analysis. The transcription consists of the start time (msec), the end time (msec), the speaker and the phoneme. The definition of the phonemes is based on that used in the project of Spontaneous speech corpus of Japanese [13]. Special symbols <cl> and ♯ designate the end of the closure in plosive or affricative sound and the start of an utterance unit, respectively.

Table 1: *Example of phonetic transcription*

483.8700	484.6700	A:	♯
484.6700	484.7550	A:	u
484.7550	484.8150	A:	<cl>
484.8150	484.8950	A:	c
484.8950	484.9450	A:	u
484.9450	485.0175	A:	n
485.0175	485.1475	A:	o
485.1475	485.2500	A:	m
485.2500	485.4500	A:	i
485.4500	485.5175	A:	y
485.5175	485.8975	A:	a

The end time differences of the utterances between the speakers are illustrated in Figure 1. Friedman's test confirmed statistically significant differences ($p < .001$). Shaffe's multiple comparison showed that the differences between the zero- and 500-msec delay groups were statistically significant.

The start time differences of the utterances between the speakers are shown in Figure 2. Friedman's test confirmed that they are not statistically significant ($p = .07$).

Figure 3 schematically illustrates the assessment parameters for the speaker efforts to adjust the utterance timing. $Ediff_{i-1}$ and $ediff_i$ are the end time differences of the utterances between the speakers. $Sdiff_i$ is the start time of the utterances between the speakers. The difference between $ediff_{i-1}$ and $sdiff_i$ signifies the speakers' collaborative effort to adjust the timing using pauses. That between $sdiff_i$ and $ediff_i$ indicates their effort to adjust the timing using utterances, that is, changing the utterance speed for their interactants.

To investigate how the speakers expended effort for synchronization, we classified the strategy patterns to adjust the timing of the utterances into (1) to (4) depicted in Figure 4. The vertical solid line shows a zero time point, where the speakers started or ended the utterances exactly at the same time. The vertical dotted line is the current point that shows when one

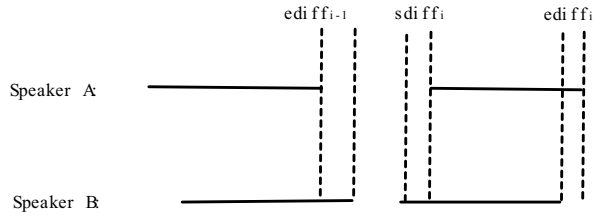

Figure 3: *Relation of timing of utterances between speakers*

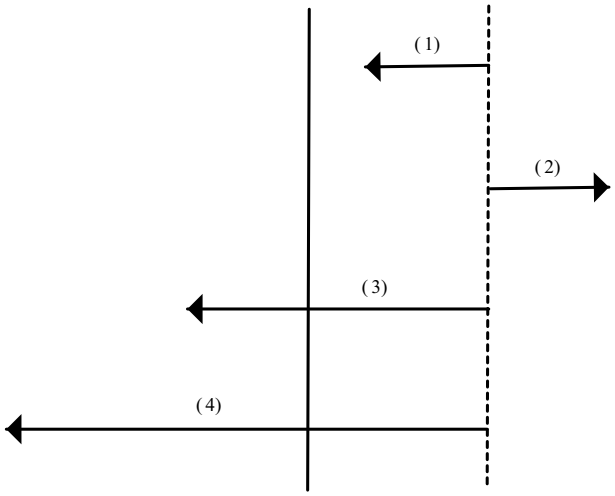

Figure 4: *Patterns of strategies for adjusting timing of utterances*

speaker's utterance lagged behind another's. Pattern (1) indicates that the speakers collaboratively adjusted their timing to reduce delay. Pattern (2) signifies cases where they failed to lessen the delay. Patterns (3) and (4) show the cases where the order of the speakers changed resulting from that one speaker started late or the other started early. Their difference lies in that the former succeeded to reduce the delay but the latter failed.

3.1. Adjusting timing using utterances

Table 2 demonstrates the frequency of the patterns to adjust the timing using utterances. χ^2 analysis does not support the independence of the cells ($p < .01$). The adjusted residual result in Table 3 indicates that the cells of (2) and (4) in the 0-msec delay condition, and those of (2) and (3) in the 500-msec delay condition are statistically larger than expected value.

3.2. Adjusting timing using pauses

Table 4 illustrates the frequency of the patterns to adjust the timing using pauses. χ^2 analysis does not maintain the independence of the cells ($p < .01$). The adjusted residual result in Table 5 shows that the cells of (2) and (3) in the 0- and 500-msec delay conditions. are statistically larger than expected value.

3.3. Sequence of patterns

Table 6 shows the sequence frequency of the patterns. χ^2 analysis does not support the independence of the cells ($p < .05$). The adjusted residual result in Table 6 shows that the cells of (1-3)-(2-4) and (2-4)-(1-3) in the 0- and 500-msec delay conditions are statistically larger than expected value.

Table 2: *Frequency of strategies for adjusting timing using utterances*

	(u-1)	(u-2)	(u-3)	(u-4)
0 msec	53	23	36	36
250 msec	49	29	26	18
500 msec	45	40	16	19

Table 3: *Adjusted residual of χ^2 analysis of Table 2*

	(u-1)	(u-2)	(u-3)	(u-4)
0 msec	-0.59960022	-2.9279990	1.6696050	2.2197865
250 msec	0.67956423	0.0567259	0.4368627	-1.3540529
500 msec	-0.05224449	3.0215000	-2.1942686	-0.9736492

4. Discussion

4.1. Adjusting the timing of speaking

For the participants, simultaneous recitation means that they starts their utterances at the same time. This does not mean that they can finish their utterances at any time. The time of finishing their utterances has an indirect effect on when they can start. Even though the effect is indirect, as Tables 2 and 3 shows, the participants in any conditions made efforts to cut down the delay. However, the speakers in the 500-msec delay condition could not reduce the delay because they spoke slowly to adapt to the partner's delayed speech. In the case where the order of the speakers changed, the speakers in the 0-msec delay condition failed to correct the difference of their utterance end times. The reason might be attributed to that changing the speed of speech is not good enough to handle such big time differences as those caused by changing the order of the speakers.

The analysis of Tables 4 and 5 illustrates that the speakers under the 500-msec condition tactfully utilized pauses to adjust the timing by changing the order of speaking and were cautious about not increasing the delay, while the speakers under the 0-msec condition did not use the pauses adequately to simultaneously start their utterances. Tables 6 and 7 also shows the tendency that the speakers in the 500-msec condition reduced and did not increase the delay using the pauses compared to those in the 0-msec condition.

4.2. Implication

The results obtained using the experiments are under conditions where the speakers were requested to recite simultaneously in auditory delayed situations, and hence seem to be limited to these conditions. However, the findings here shed light on a potential of speakers to adjust the timing of their utterances, which can be used for deepening the understanding of dynamics of conversation and for building applications.

Research on the system of turn-taking such as [10, 14] elucidated the signal and the regularities observed in orderly change of speakers, but did not provide explanation on how speakers influence each other like adoption of speech rate [4]. Suppose that virtual agents or robots are talking to young children or senior citizens who might have difficulty understanding if the agents or robots speak too slowly or too fast. In such cases only smooth change of turns cannot realize comfortable communication. The agents or robots need to coordinate the

Table 4: *Frequency of strategies for adjusting timing using pauses*

	(p-1)	(p-2)	(p-3)	(p-4)
0 msec	23	39	22	41
250 msec	26	30	29	37
500 msec	18	16	42	44

Table 5: *Adjusted residual of χ^2 analysis of Table 4*

	(p-1)	(p-2)	(p-3)	(p-4)
0 msec	0.05127625	2.6237789	-2.4502154	-0.1293312
250 msec	1.06919801	0.4580533	-0.4879874	-0.8364069
500 msec	-1.12549885	-3.1106215	2.9653411	0.9705822

Table 6: *Frequency of sequences of strategies for adjusting timing*

	(1,3)-(1,3)	(1,3)-(2,4)	(2,4)-(1,3)	(2,4)-(2,4)
0 msec	25	64	20	16
250 msec	28	47	27	20
500 msec	25	36	35	24

Table 7: *Adjusted residual of χ^2 analysis of Table 7*

	(1,3)-(1,3)	(1,3)-(2,4)	(2,4)-(1,3)	(2,4)-(2,4)
0 msec	-0.4218344	3.131671	-2.09668880	-1.32123458
250 msec	0.5609026	-0.422076	-0.06885957	0.01632884
500 msec	-0.1371101	-2.739880	2.18730352	1.31836348

speech with their partners, which can be attained by changing the speed of utterances and the length of pauses.

5. Conclusion

This paper explored a mechanism for synchronizing speech using auditory delayed experiments where the speakers simultaneously recited. Statistical analysis demonstrated that the speakers adjusted the timing of speech by changing the speed of utterances and the length of pauses.

Since speech affects (or be affected by) bodily movements, adjusting behavior of speech timing will have some relation with bodily movements. For example, measuring the timing for such non-speech intervals as pauses is difficult, and hence speakers might use bodily movements for this task, which constitutes the target of our next research.

6. Acknowledgements

The authors would like to Dr. Dean Hay of the graduate school of Interdisciplinary Information Studies, The University of Tokyo for his help and advice in conducting the experiments.

7. References

[1] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R., "The HCRC Map Task Corpus," Language and Speech, 34, 351–366, 1991.

[2] Austin, J. L., "How to Do Things with Words," Oxford University Press, 1962.

[3] Bickmore, T. and Cassell, J., "Social Dialogue with Embodied Conversational Agents," In J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems, Kluwer Academic, (in press).

[4] Bosshardt, H.-G., Sappok, C., Knipschild, M. and Hölscher, C., "Spontaneous Imitation of Fundamental Frequency and Speech Rate by Nonstutterers and Stutterers," Journal of Psycholinguistic Research, 26(4), 425–448, 1997.

[5] Branigan, H. P., Pickering, M. J. and Cleland, A. A., "Syntactic Coordination in Dialogue," Cognition 75, 13-25, 2000.

[6] Brennan, S. E. and Clark, H. H., "Conceptual pacts and lexical choice in conversation," Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1482-1493, 1996.

[7] Clark, H. H. and Fox Tree, J. E., "Using uh and um in spontaneous speech," Cognition, 84, 73-111, 2002.

[8] Clark, H. H. and Krych, M. A., "Speaking while monitoring addressees for understanding," Journal of Memory and Language, 50(1), 62-81, 2004.

[9] Couper-Kuhlen, E., "English speech rhythm: Form and function in everyday verbal interaction," John Benjamin Publishing Company, 1998.

[10] Duncan, Jr., S., "Some signals and rules for taking speaking turns in conversations," Journal of Personality and Social Psychology, 23, 283–292, 1972.

[11] Traum, D. R., Swartout, W., Gratch, J. and Marsella, S., "A Virtual Human Dialogue Model for Non-team Interaction," in Recent Trends in Discourse and Dialogue Springer, Dybkjaer, L. and Minker, W. (Eds.) 45–67, 2008.

[12] Den, Y., Ishizaki, M. and Fukashiro, S., "Conforming behavior of speech and bodily movement: Analysis of the sway in auditory delayed conditions," The proceedings of the 24th Annual Meeting of the Japanese Cognitive Science Society, 426–429, 2007 (in Japanese).

[13] Maekawa, K., Koiso, H., Furui, S. and Isahara, H., "Spontaneous speech corpus of Japanese," Proceedings of the Second International Conference of Language Resources and Evaluation (LREC2000), 947–952, 2000.

[14] Sacks, H., Schegloff, E. A. and Jefferson, G., "A simplest systematics for the organization of turn-taking for conversation," Language, 50, 696–735, 1974.

[15] Searle, J. R., "Speech Acts: An Essay in the Philosophy of Language," Cambridge University Press, 1969.

[16] Shockley, K., Santanna, M.-V. and Fowler, C. A., "Mutual interpersonal postural constraints are involved in cooperative conversation," Journal of Experimental Psychology: Human Perception and Performance, 29, 326–332, 2003.