Cross-Modal Interaction in Graphical Communication

Ichiro Umata (umata@atr.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan
Yasuhiro Katagiri (katagiri@atr.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan

Abstract

Cross-modal interaction in graphical communication is observed in collaborative problem-solving settings. Graphical communications, such as dialogues using maps, drawings, or pictures, provide people with two independent modalities: speech and drawing. Although the amount of drawing/self-speech overlap is strongly affected by activity-dependent constraints imposed by the task, the amount of drawing/partner's speech overlap is affected only weakly by these constraints. However, they do affect the function of the utterances in the case of drawing/partner's speech overlap. These results show that activity-level constraints affect the way speech coordinates drawing activities in cross-modal interaction. Furthermore, it suggests that turn-taking in multimodal communication requires general analyses integrating the functions of different modalities.

Introduction

Every joint activity requires coordination among its participants. When a band plays a piece, each member has to work on the same key, keep the same rhythm, and start and end at the same time (Clark (1996)). Some of these coordinating acts can be done across different modalities. In the case of music, a soloist can signal the end of her inprovisation not only with a phrase suggesting the solo's end, but also with eye-contact.

Communication is also a joint activity, and participants must coordinate with each other. One outstanding coordination principle in conversation is sequential turn-taking in speech channels. Several studies have been carried out on speech turn coordination, and some of them analyze cross-modal interaction between speech and nonverbal behaviors such as gaze and posture (Argyle et al. (1976), Kendon (1967)). In this paper, we investigate the interaction between speech and drawing, another powerful communication medium.

Turn-taking in speech involves a wide variety of factors such as sociological principles, the limitations of human cognitive capacity, and so on. One potentially strong factor for sequential turns in speech is the resource characteristics of media: speech media affords only one person's speech sounds at a time. Sacks et al. (1974) regard verbal turns as an economic resource, distributed to conversation participants according to turn organization rules. According to them, one of the main effects of these turn organization rules is the sequentiality of utterances. They observe that one party talks at a time in most cases.

Drawing, on the contrary, has quite different characteristics from speech. First, drawing is persistent whereas speech is not. Drawing remains unless erased, whereas speech dissipates right after it occurs. A drawing can be understood much later than when it is actually drawn, whereas speech must occur in real time. Second, drawing has a much wider bandwidth than speech. Two or more drawing operations can occur at the same time without interfering with each other, whereas simultaneous utterances are hard to understand. These resource characteristics allow for simultaneous drawing. There have been several studies on drawing interaction in the Human Computer Interaction field in the context of computer-supported collaborative work. Some researchers are optimimistic about the possibilities of simultaneous drawing (Stefik et al. (1987), Whittaker et al. (1991)), though others are not (Tatar et al. (1991)).

To approach this problem, Umata et al. (2003) have introduced yet another view based on the activity-dependent constraints imposed by the task performed in the interaction. The analyses show that sequential structure is mandatory in drawing either when the drawing reflects the dependency among the information to be expressed or when the drawing process itself reflects the proceedings of a target event. Further analyses show that speech interaction, which is already restricted by the resource characteristics of media, is not affected by activity-dependent constraints (Umata et al. (2004)).

The relation between drawing and speech modalities is, however, still not quite clear. Takeoka et al. (2003) analyzed face-to-face graphical communication and found that both utterances without drawings and utterances followed by the speaker's drawings behave similary in turn-holding function. They also show that longer silences are allowed while drawing is taking place. These results suggest that turns in communication can be maintained across speech and drawing modalities. This is also supported by the finding that drawing/self-speech overlap is much more frequent than drawing/partner's speech overlap (Umata et al. (2004)). The assumption of continuous turns across modalities is appealing from the viewpoint of modal integration: speech and graphic modalities decribe their target not just independently but also jointly, with linguistic phrases describing the target via graphics (Umata et al. (2000)).

In the following part of this paper, we analyze interaction across these two modalities, focusing on drawing-speech overlap. The results show that the activity-dependent constraints strongly affect the amount of drawing/self-speech overlap, whereas they only weakly affect the amount of drawing/partner's speech overlap. These constraints, however, do affect how their drawing activities are coordinated verbally. We argue that activity-level constraints affect not only drawing-drawing interaction organization but also cross-modal interaction organization.

Drawing Turns and Speech Turns

As we have seen in the previous section, the sequentiality of speech turns has been attributed to the resource characteristics of speech, namely non-persistence and restricted bandwidth. The assumption is that we cannot comprehend two spoken utterances at the same time because of the bandwidth limitation, while we cannot delay comprehending one utterance until later because of the non-persistent characteristic. Drawing, on the contrary, functions quite differently in regard to these assumptions, and it may have potential for parallel turn organization. There have been seemingly contradictory observations of drawing turn organization; one is that drawing turns can be parallel, and the other is that they cannot be parallel. Umata et al. (2003) suggested that there is yet another kind of constraint based on the activities people are engaged in. According to this view, sequential structure is mandatory in drawing in some cases but not in others.

Sequentiality Constraints

1. Drawing interaction occurs in sequential turns under either of the following conditions:
   (a) Information Dependency Condition: When there is a dependency among the information to be expressed by drawing;
   (b) Event Alignment Condition: When drawing operations themselves are used as expressions of the proceedings of target events.
2. Sequential turns are not mandatory in drawing activities when neither condition holds (and when persistence and certain bandwidths of drawing are provided).

The rationale for the information dependency condition is the intuition that when one piece of information depends on another, the grounding of the former piece of information is more efficient *after* the grounding of the latter has been completed. This should be the case whether a particular speaker is explaining the logical dependency in question to her partners or all participants are following the logical steps together.

Event alignment is a strategy for expressing the unfolding of an event dynamically, using the process of drawing itself as a representation. For example, when you are reporting on how you spent a day in a town by using a map, you might draw a line that shows the route you actually took on the map. In doing so, you are aligning the drawing event with the walking event to express the latter dynamically. Our hypothesis is that simultaneous drawing is unlikely while this strategy of event alignment is employed. Under this condition, the movement or process of drawing is the main carrier of information. The trace of drawing has only a subsidiary informational role. Thus, in this particular use of drawing, its persistency is largely irrelevant. The message must be comprehended and grounded in real time, and the bandwidth afforded by the drawing surface becomes irrelevant. This requirement effectively prohibits the occurrence of any other simultaneous drawing.

An analysis on the corpus gathered from collaborative problem-solving tasks demonstrates that these two activity-dependent constraints can override the resource characteristics of the drawing media, thereby enforcing a sequential turn organization similar to those observed in verbal interactions (Umata et al. (2003)).

These activity-dependent constraints, however, do not affect the speech turn organization that is already affected by resource characteristics. The amount of simultaneous speech shows no difference among different task conditions (Umata et al. (2004)).

In the following part of this paper, we will look into the details of cross-modal overlap, based on the analysis of collaborative problem-solving task data gathered by Umata et al. (2003). We will compare the speech turn organization patterns in different task settings to see whether activity-dependent constraints affect the amount of drawing-speech overlap.

Method

An experiment in which subjects were asked to communicate graphically was conducted to examine the effect of the two factors presented above on their interaction organization. In these experiments, 24 pairs of subjects were asked to work collaboratively on four problem-solving tasks using virtual whiteboards.

Experimental Setting

In the experiments reported here, two subjects collaboratively worked on four different problem-solving tasks. All of the subjects were recruited from local universities and paid a small honorarium for their participation. The subjects were seated in separate, soundproof rooms and worked together in pairs using a shared virtual whiteboard (50 inches) and a full duplex audio connection. The subjects were video-taped during the experiment. They also wore cap-like eye-tracking devices that provided data indicating their eye-gaze positions. The order in which the tasks were presented was balanced between the 24 pairs so that the presentation order would not have an affect on the results. The time limit for each task was six minutes.

At the start of each task, an initial diagram was shown on the subjects' shared whiteboard and the subjects were then free to speak to one another and to draw and erase on the whiteboard. The only limitation to this drawing activity was that they could not erase or occlude the initial diagram. All drawing activity on the whiteboard was performed with a hand-held stylus directly onto the screen, and any writing or erasing by one participant appeared simultaneously on the whiteboard in the partner's room. The stylus controlled the position of the mouse pointer and, when not drawing, the positions of both subjects' mouse pointers were displayed on the shared whiteboard.

Tasks

Deduction Task with an Event Answer (1e) A logical reasoning problem with a correct answer. The problem asks the subjects to describe the arrangement of people around a table and the order in which the people sit down. This seating arrangement and order must satisfy some restrictions (e.g., "The fifth person to sit is located on the left-hand side of person B."). A circle representing a round table was shown on the whiteboard at the start of the task. This task has strong informational dependency and strong event alignment.

Deduction Task with a State Answer (1s) A logical reasoning problem with a correct answer asking that the subjects design a seating arrangement satisfying some restrictions (e.g., "S cannot sit next to M."). A circle representing a round table was shown on the whiteboard at the start of the task. This task has strong informational dependency and loose event alignment.

Design Task with an Event Answer (2e) A task with an open-ended answer, asking subjects to make an excursion itinerary based on a given town map. A complete town map was shown on the whiteboard at the start of the task. This task has weak informational dependency and strong event alignment[1].

Design Task with a State Answer (2s) A task with an open-ended answer, asking the subjects to design a town layout to their own liking. An incomplete town map was shown on the whiteboard at the start of the task. This task has weak informational dependency and loose event alignment.

Data

During each task, all drawing, erasing, and mouse movements by each subject were recorded in a data file. Using this data, the amount of simultaneous drawing was calculated as the total time spent drawing simultaneously as a percentage of the total time either subject spent drawing (i.e., the sum of the time intervals in which both subjects drew simultaneously divided by the sum of the time intervals in which at least one of the pair drew on the whiteboard). Speech was recorded with video-data and labeled by hand. As with the drawing data, the amount of simultaneous speech was calculated as the total time spent talking simultaneously as a percentage of the total time either subject talked.

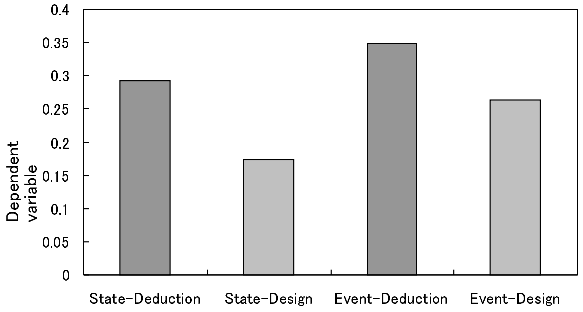

Figure 1: Proportion of drawing/self-speech overlaps

Analysis 1

Drawing/Self-Speech Overlap

As shown in Figure 1, the proportion of drawing/self-speech overlap time to total drawing time was the smallest in the design state (2s) condition. This data was entered into a 2 x 2 Analysis of Variance (ANOVA). Both problem type (deduction and design) and solution type (state and event) were treated as within-subject factors. Analysis revealed a main effect of problem type $F(1,47)=24.968$, $p<.001$ and solution type $F(1,47)=21.783$, $p<.001$ and showed no interaction $Fs < 1$.

Thus, it was shown that the proportion of drawing/self-speech overlap is smaller when the task has either weaker informational dependency or weaker event alignment, or both.

Drawing/Partner's Speech Overlap

As shown in Figure 2, the proportion of drawing/partner's speech overlap time to total drawing time demonstrated a significant, but smaller, difference in each condition compared to the case of self overlap. This data was entered into a 2 x 2 ANOVA. Both problem type (deduction and design) and solution type (state and event) were treated as within-subject factors. Analysis showed a simple main effect of solution type $F(1,47)=4.484$, $p=.04$. No effect was found for the problem type, and analysis showed no interaction $Fs < 1$.

The analysis showed that the proportion of drawing/partner's speech overlap is only weakly affected by the event alignment condition.

Discussion for Analysis 1

The amount of drawing/self-speech overlap is smaller when the task has either weaker informational dependency or weaker event alignment, or both. The activity-dependent constraints work on self-cross-modal overlap

[1] Note that these categories are relative rather than absolute. For example, (2e) also has informational dependency to a certain extent in that each path has to start from the icon of the previous place they decided to visit. However, they can choose the next destination freely. Thus informational dependency is much weaker than in the cases of the seat arrangement tasks where one decision significantly narrows down the subsequent alternatives; e.g., seating a person *M* in a certain position means only *S* or *O* can sit right next to *P*, and so on.

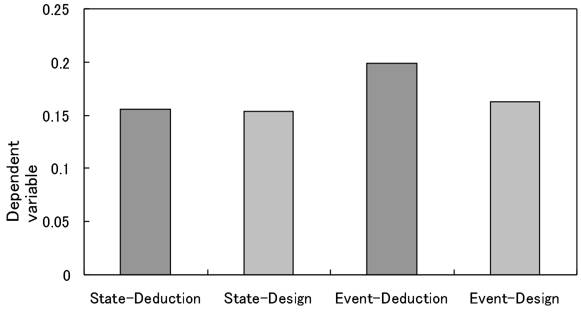

Figure 2: Proportion of drawing/partner's speech overlaps

in the opposite way of simultaneous drawing: the amount of simultaneous drawing is smaller when the task has stronger information dependency or weaker event alignment, or both.

This result seems quite reasonable if we consider the way people coordinate their drawing activities verbally. Whittaker et al. (1991) observed verbal coordination of drawing activities through the examination of shared whiteboard communication *with* and *without* the addition of a speech channel. They found that permanent media such as a whiteboard provides users with space for constructing shared data structures around which they can organize their activity. With the addition of a speech channel, people used the whiteboards to construct shared data structures that made up the CONTENT of the communication, while speech was used for coordinating the PROCESS of communication.

As observed in Umata et al. (2004), utterances coordinating drawing activities are also commonly found in our tasks. Figure 3 is a snapshot from the deductive state task (1s). Subjects *A* and *B* have just agreed to fix *M*'s seat first, and *A* suggests "*M*'s seat should be ... here, right?" while drawing the sign M. Then, *B* gives verbal acknowledgement, "Yes." Here, *A*'s utterance serves as a signal for his drawing activity.

Such signal utterances typically preceed drawings, and drawings follow, overlapping them. Signal utterances are expected to occur more often when people feel a stronger need to coordinate their drawing activities; i.e., in cases where activity-level constraints require sequential drawing turns. As expected, drawing/self-speech overlap is most frequent when the task has strong informational dependency or tight event alignment, or both.

There are two other possible explanations for the result. The first is that drawers have to give more verbal explainations of what they are doing as the task increases in difficulty. This does not seem to be the case, though. First, those signal utterances are usually quite simple and short: e.g., "*M* is here," "Station," etc. Second, their drawings are generally simple and easy to understand even in the tasks with stronger constraints. In the seat arrangement tasks ((1e), (1s)), each icon is an alphabetic letter standing for a person. Its position on the table icon simply shows where the person has to be seated, and the sequence of letters beside the table icon means the order of the seating. In the case of the excursion itinerary task (2e), the drawings were mainly route icons and labels showing time of arrival/departure and so on. The meaning of each drawing is also clear to the partner in this case. On the other hand, some icons can be unintelligible to the partners in the case of the town layout task (2s): a box can mean a building icon, a station icon, or anything else. Actually, people sometimes had to ask their partners for more clarification in (2s). Thus, the utterances about what they are drawing are likely to be just signals rather than detailed explanation of their drawing.

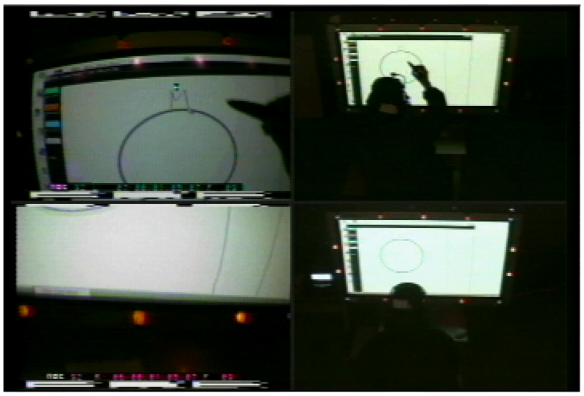

Figure 3: Sequential drawing interaction coordinated verbally (1)

The second possible explanation is that simultaneous drawing and cross-modal overlap are affected not by the activity-level constraint but by the symbolic status of the drawing. That is, the drawing requires sequential drawing turn organization in (1s), (1e) and (2e) because they are not just a set of icons but rather a language-like symbolic system. This is also unlikely, since the drawings are almost equally simple throughout the tasks, as described above. It is possible, though, that more complicated symbolic systems require sequential turns and that it is difficult to separate the effect of the activity-level constraint and that of symbolic constraction. More work is required to illuminate the detailed mechanism underlining sequential drawing turn organization.

The activity-level constraints have a much weaker effect on the amount of drawing/partner's speech overlap. Because people cannot precisely predict when and where their partner will start drawing, verbal coordination of drawing activities typically takes the form of signal utterances. This may be why these constraints did not impact strongly on the amount of drawing/partner's speech overlap.

Another possible explanation is that turns in graphical communication tend to be maintained across speech and drawing modalities. Drawing/self-speech overlap is much more common than partner's speech overlap

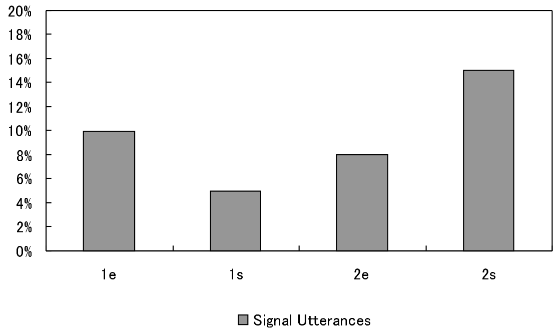

Figure 4: Frequencies of verbal signals for drawing

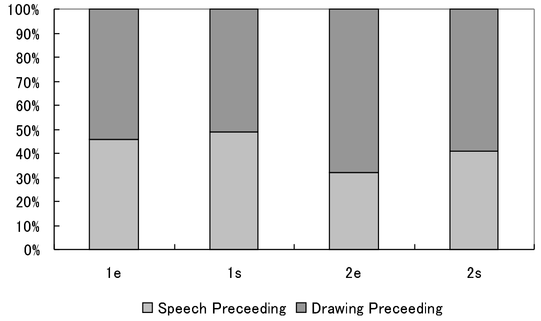

Figure 5: Frequencies of utterance preceeding overlap

(Umata et al. (2004)). The effect of activity-level constraints is much weaker, perhaps because the cross-modal turn organization already blocks speech overlap by partners.

Analysis 2

It was shown that the activity-dependent constraints affect the amount of drawing/partner's speech overlap only weakly, whereas they strongly affect the amount of drawing/self-speech overlap. In this section, we analyze drawing/partner's speech overlap in more detail to determine whether there are any differences among task conditions.

The drawing occurences analyzed above were all recorded as the time duration that the pen is touching the screen. Some drawing activities are divided into segments that are too small under this method. For example, some subjects drew many dots or lines to give colors to some icons. It is unreasonable to divide such an activity into many drawing occurences when we perform closer analysis on each overlapping case of drawing and speech modalities. The drawing occurences within 400 msec gaps are regarded as a *drawing unit* for the analysis below, in the same way as when we divide speech into utterance units. One member of each of the 24 dyads tested was randomly selected for the following analyses.

Verbal Signals for Drawings

The frequencies of verbal signals in all drawing/partner's speech overlap were compared among different task conditions. The analysis showed significantly different proportions among conditions ($\chi^2{}_{(3)} = 13.775, p < .003$). More concretely, verbal signals in drawing/partner's speech overlap are most frequent in the design state condition (2s), as shown in Figure 4 (adjusted residual: (1e) $= -1.2$, (1s) $= -8.7$, (2e) $=$ -5.4, (2s) $= 15.3$). The design state condition has fewer verbal signals for drawing overall, so their high frequency in drawing/partner's speech overlap is rather outstanding.

Other Findings: Drawing Preceeded Overlaps

We also compared the frequencies of drawing preceeding overlap in drawing/partner's speech overlap among different task conditions. The analysis shows significantly different proportions between deduction conditions (1e, 1s) and design (2e, 2s) conditions ($\chi^2{}_{(3)} = 7.740, p <$ .005). More concretely, the design conditions have fewer drawing preceeding overlap than the deduction conditions, as shown in Figure 5 (adjusted residual: (1e, 1s) $= -19.2$, (2e, 2s) $= 19.2$). People start drawing while their partners are speaking more often in the design condition than the deduction condition.

Discussions for Analysis 2

Drawing/partner's speech overlap includes more verbal signals for drawings in the design state condition (2s) than in any other condition. This reflects the parallel interaction style of drawing in (2s). While verbal signals serve to maintain sequential drawing interaction in the case with stronger activity-level constraints, these signals often serve to coordinate parallel drawing activities when they occur in (2s). Verbal signals also overlap the partner's drawings in some of these cases. Figure 6 shows one such case. Subjects *A* and *B* agreed to divide the design task into two sub-tasks, the design of a station plaza and that of a park. Then, *A* said "Station," and *B* said "I'll make the forest," before starting their respective drawing activities. Here, they verbally coordinated their simultaneous drawing activity, and their verbal signals overlap their partner's drawings.

Drawing preceeding overlap is more frequent in the design condition (1) than in the deduction condition (2). This means only the information dependency constraint affected the frequency of speech preceeding overlaps. Although we cannot give any clear explanation for this phenomenon, we assume this result reflects the different characteristics of these two activity-dependent constraints. The information dependency constraint has a more general nature across modalities: when one piece of information depends on another, the grounding of the former piece of information is more efficient after the grounding of the latter has been completed. On the contrary, event alignment is rather drawing-modality-specific: the drawing process reflects the process of the described event. In this sense, drawing activities are less dependent on the information given in speech modalities

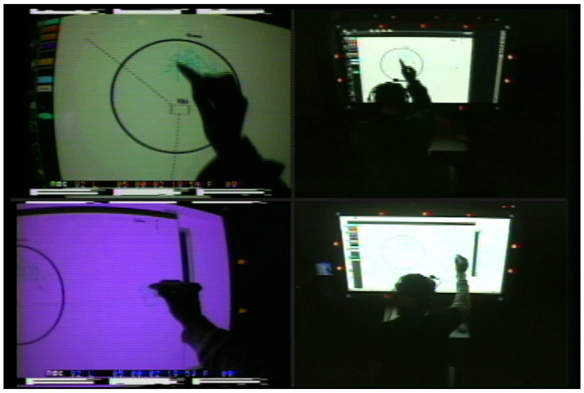

Figure 6: Parallel drawing interaction coordinated verbally

than in cases with strong information dependency. However, the mechanism causing this phenomenon remains unclear. More work is required to demonstrate how the two modalities interact.

Conclusions

Based on the data of collaborative task solving settings, we have analyzed cross-modal interaction in graphical communication. We found that the amount of drawing/self-speech overlap is strongly affected by the activity-dependent constraints, while the amount of drawing/partner's speech overlap is affected only weakly by these constraints.

There are, however, significant differences in the function of the utterances in the case of drawing/partner's speech overlap. Drawing/partner's speech overlap includes more signal utterances for drawing when the activity-level constraints are weaker. This result reflects the parallel interaction style of drawing under weak activity-level constraints.

The precedence of drawing/partner's speech overlap is also affected by the information dependency constraint. Although it is likely that the modality-general nature of this constraint plays a significant role, the mechanism of this phenomenon is still not clear.

These findings indicate that the activity-level constraints affect the way speech coordinates drawing activities in cross-modal interaction and suggest that interaction organization in multimodal communication is a complex phenomenon that requires general analyses integrating the functions of different modalities.

Acknowledgments

This research was supported in part by The National Institute of Information and Communications Technology (NICT) of Japan. We would like to thank the anonymous referees for their insightful comments. Thanks also must go to Takugo Fukaya for his intellectual support of this work.

References

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Brennan, S. E. (1990). Seeking and providing evidence for mutual understanding. Ph.D. dissertation, Stanford University.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259–294.

Condon, W. S. (1971). Speech and body motion synchrony of the speaker-hearer. In D. L. Horton and J. J. Jenkins (Eds.), *Perception of Language*. Columbus, Ohio: Merill, 150–173

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32*, 1–25.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50*, 696–735.

Stefik, M., Foster, G., Bobrow, D., Kahn, K., Lanning, S., and Suchman, L. (1987). Beyond the chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of the ACM, 30 (1)*, 32–47.

Takeoka, A., Shimojima A., and Katagiri, Y. (2003). Turn-taking in graphical communication: An exploratory study. In *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*.

Tatar, D., Foster, G., and Bobrow, D. (1991). Design for conversation: Lessons from cognoter. *International Journal of Man-Machine Studies, 34(2)*, 185–210.

Traum, D. (1994). A computational theory of grounding in natural language conversation. Ph.D. dissertation, University of Rochester.

Umata, I., Shimojima, A., and Katagiri, Y. (2000). Talking through graphics: An empirical study of the sequential integration of modalities. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 529–534.

Umata, I., Shimojima, A., Katagiri, Y., and Swoboda, N. (2003). Graphical turns in multimodal communication. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (CD-ROM).

Umata, I., Shimojima, A., Katagiri, Y. (2004). Speech and graphical interaction in multimodal communication. To appear in *Proceedings of Diagrams 2004*.

Whittaker, S., Brennan, S., and Clark, H. (1991). Coordinating activity: An analysis of computer supported co-operative work. In *Proceedings of CHI'91 Human Factors in Computing Systems*, 361–367, New York: ACM Press.