AAAI Fall Symposium on Embodied Language and Action,
Massachusetts Institute of Technology, November, 1995, 102-108.

Computational Characteristics of Multimodal Dialogue

Kris R. Thórisson
The Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
kris@media.mit.edu

Abstract

This paper examines some of the general characteristics of real-time multimodal, face-to-face interaction that set it apart from other issues in AI and human-computer interaction. It argues that a real-time multimodal system needs to be designed in layers and include both reactive and reflective behaviors. It presents a layered feedback-loop model of face-to-face dialogue and shows how contextual analysis of the function of multimodal acts is linked with feedback generation and interpretation. A brief description is given of a modular architecture called Ymir, based on the proposed model, for modeling psychosocial dialogue skills.

Introduction

The work described in this paper is motivated by computer interfaces that depend on a metaphor of human face-to-face communication [Thórisson 1994, 1993, Laurel 1990, Bolt 1987, Bolt 1985]. Such interfaces deploy an artificial agent that responds to human multi-modality and is capable of multimodal behavior and actions in a limited domain such as graphics generation or information retrieval [Koons et al. 1993, Laurel 1990]. In order for the multimodal interface agent metaphor to work, the mechanisms controlling the on-screen agent have to capture elements that are critical to the structure of multimodal dialogue, such as gestural meaning, body language, turn-taking, etc., and integrate these in a comprehensive way.

The current work deals not with an isolated, single process or problem within face-to-face interaction but the larger picture of bridging between input and output to close the full loop of multimodal interaction between the human and machine. The premise behind this approach is that *reciprocity* is key in multimodal dialogue. To address the problems encountered in "full-duplex" multimodal interaction, an architecture of multimodal dialogue skills is being developed that bridges between input analysis and output generation and serves as a testbed for multimodal agents. A number of general observations underlie the approach which should be useful for anyone working on real-time multimodal systems. This paper addresses the many premises and assumptions, arguing particular points deemed important to the creation of such systems. The multimodal architecure being developed is described briefly at the end of the paper

Challenges of Real-Time Multimodal Interaction

From a computational perspective, many features set real-time face-to-face interaction apart from other topics in human-computer interaction and artificial intelligence. For the current purposes, these may be identified as:

1. Incremental interpretation,
2. multiple data types,
3. seamlessness,
4. temporal constraints, and
5. multi-layered input analysis and response generation.

1. Multimodal interpretation is not done "batch-style:" There are no points in an interaction where a full multimodal act or a

whole sentence is output by one participant before being received by another and interpreted as a whole. Interpretation happens in parallel with multimodal output generation.

2. Multimodal interaction contains many data types, in the traditional computer science sense, as any quick glance at the main communicative modes will show: Gestures [McNeill 1992, Goodwin 1986, Ekman 1979, Ekman & Friesen 1969] provide spatial and relational information, speech [Allen 1987, Goodwin 1981] provides semantic and prosodic information [Pierrehumbert & Hirschberg 1990], gaze [Kleinke 1986, Argyle et al. 1974, Kahneman 1973], head and body provide directional[1] data related to attention and the dialogue process.

3. When interacting with each other, people generally are not aware of the fact that interjecting for example an iconic gesture into the discourse constitutes a different kind of information than a deictic one, and they don't particularly notice the mechanism by which they take turns speaking. The various data types encountered in face-to-face dialogue have to be recognized automatically to allow us to communicate efficiently with a multimodal agent.

4. The structure of dialogue requires that participants agree on a common speed of exchange [Goodwin 1981]. If the rhythm of an interaction is violated, the violating participant should make this clear to others so that they can adjust to the change. Actions also have to be produced in a timely manner: a listener's glance, for example, in the direction that the speaker pointed has a different meaning if it happens 10 seconds after the deictic gesture was made.

5. In discourse, responses in one mode may overlap those of another mode in time, and constitute different information [McNeill 1992, Goodwin 1981]. The layers can contain anything from very short responses like glances and back channels [Yngve 1970], to tasks with longer time spans, such as whole utterances and topic continuity generation. In order for purposeful conversation to work, reactive and reflective[2] responses have to co-exist to provide for adequate behavior of an agent.

When trying to incorporate the above principles into the design of artificial agents, it becomes apparent that certain additional characteristics of the human interpretive processes and quality of "input data" have to be taken into consideration:

1. Interpretation is fallible,
2. there are both deficiencies and redundancies in input data,
3. sensory data is collected to allow an agent to produce action or inaction,
4. behavior is based on data from multiple sources, both internal and external, including dialogue state, body language, etc., and
5. behavior is eventually always produced, no matter what data is available.

Because of inaccuracies in the information delivery of humans, among other things, the first assumption will hold no matter how powerful our interpreter is, whether human or artificial. This problem is worsened in artificial agents by the use of faulty sensors, occlusion when using cameras, etc. The second item points to the inevitable fact of having to deal with missing information, and, in certain cases, redundancy as a possible solution to that problem, both in interpretation and output generation. The third item reflects the purpose-directed sensory and cognitive abilities of any situated agent, and clearly points to the need of an ego-centered design when producing social behavior in machines. Item four makes it clear that in multimodal communication action can be—and perhaps most often is—taken based on more than a single piece of information. The fifth item points out that both a listener and a speaker in dialogue are expected to exhibit the necessary behaviors pace the interaction. An agent cannot therefore be solely event driven—it has to be autonomous to some extent.

[1] My thanks to Steve Whittaker for the term *directional* in this context.

[2] I use the terms *reactive* and *reflective* behaviors as a shorthand to refer to the time-scale of the behavior—i.e. the *sense-act cycle.* Generally speaking, reactive behavior can be found in the lower bound of the cycle (0-1 second range), while reflective behaviors generally range anywhere from a second to hours.

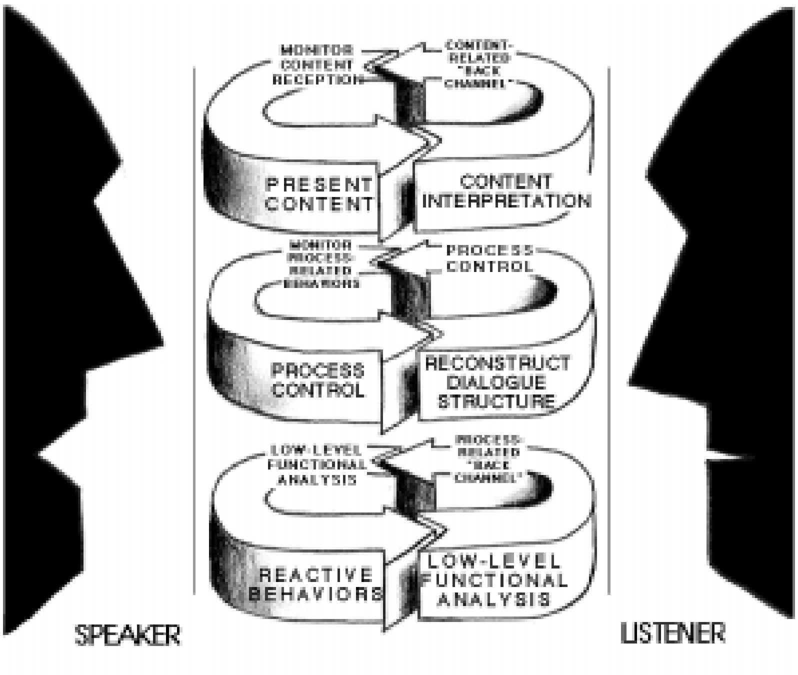
MONITOR CONTENT RECEPTION
CONTENT-RELATED "BACK CHANNEL"
PRESENT CONTENT
CONTENT INTERPRETATION
MONITOR PROCESS-RELATED BEHAVIORS
PROCESS CONTROL
PROCESS CONTROL
RECONSTRUCT DIALOGUE STRUCTURE
LOW-LEVEL FUNCTIONAL ANALYSIS
PROCESS-RELATED "BACK CHANNEL"
REACTIVE BEHAVIORS
LOW-LEVEL FUNCTIONAL ANALYSIS
SPEAKER
LISTENER

Figure 1. The proposed three-layered model of multimodal dialogue.

Multimodal Interaction as Layered Loops

The model put forth here of multimodal interaction can be characterized as a layered feedback-loop[3] model, and is intended to be prescriptive (Figure 1). The three layers in the model are based on the time-scale of actions found in face-to-face dialogue (Figure 2). At each level various sensory and action processes are active, whose type is mostly determined by the role of the participant: *speaker* or *listener*. The lowest level is concerned with behaviors that generally require recognize-act cycles shorter than 1 second. This is the *Reactive* layer. The middle layer concerns behaviors that usually are slower than 1 second. This is the *Process Control* layer. Together these two layers define the mechanisms of dialogue management, or psychosocial dialogue skills. Highly reactive actions, like looking away when you believe it's your turn to speak [Goodwin 1981] or gazing at objects mentioned to you by the speaker [Kahneman 1973], belong in the lowest layer. Direct references to the process of dialogue ("I'm trying to remember..." and "Let's see...") belong in the Process Control layer and are generated in response to the status of processes in the other layers. The scheduling of other high-level responses—e.g. those generated in response to the content of the dialogue—are also managed in this layer. The third part of this model is the *Content* layer, where the topic of the conversation is processed. This layer deserves its own discussion, and will not be dealt with here. We will now examine the dialogue management layers more closely.

Desired Characteristics of Multimodal Systems

In this section we will look at the three following claims: (1) To produce coherent behavior in real-time dialogue, *reactive* and *reflective* behaviors have to co-exist in the same system, (2) analysis of the contextual *function*[4] of speaker actions

[3] "Feedback" in this context refers to the reciprocal nature of any speaker-hearer relationship, where a participant's [P1] multimodal action [P1-1] is met by the other's [P2] *re*-action [P2-1]. This loop can be more than one level deep; a common format is the sequence [P1-1∅P2-1∅P1-2].

[4] My use of the term "function" is roughly equvalent to its use in speech act theory [Searle

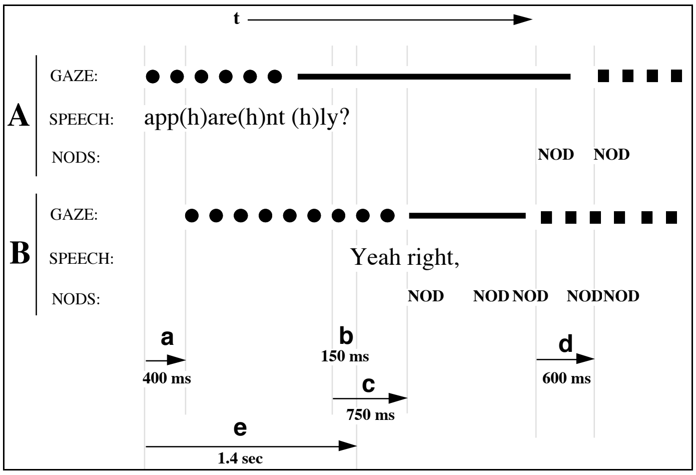

Figure 2. Transcript spanning 3 seconds of a typical two-person conversation, showing the timing of speech, gaze and head nods for each conversant (Adapted from Goodwin [1981]). "**A** brings her gaze to the recipient. **B** reacts to this by immediately bringing her own gaze to **A**. The two nod together and then ... withdraw from each other, occupying that withdrawal with a series of nods" [Goodwin 1981, p. 119]. Notice that **a**, **b**, **c** and **d** are listener reactions to speaker actions; these all happen under 1 second. **b** is a turn transition. **e** is the estimated minimum time the listener had for generating a response to the content of the speakers preceding turn.

and control of the *process* of dialogue are intimately linked through what I refer to as *functional analysis*, and (3) the information necessary for *correct and efficient content analysis* is also the necessary information for providing *correct and efficient multimodal feedback behavior.*

Combining Reactive and Reflective Behaviors for Real-Time Response

Face-to-face conversation is unique because it contains processes that span as much as five orders of magnitude of execution time, from[5] about 100 ms to minutes and hours [Thórisson 1994]. A look at the transcription in Figure 2 [Goodwin 1981] shows that in face-to-face discourse, rapid responses and more reflective ones are interwoven in a complex pattern. This kind of interaction is the basis for the dialogue management system proposed.

The Listener's Functional Analysis of Speaker Behavior: A Precursor to Content Interpretation and Feedback Generation

Low-level (basic, elementary) interpretation of a speaker's behavior should not primarily be concerned with what lexical elements can be best mapped onto the user's utterance, or whether the utterance at any point in time is grammatically correct. It should be concerned with distinctions that determine broad strokes of behavior, i.e. extracting the features that make the major distinctions of the dialogue. For example, computing answers to questions like "is this person addressing *me?"* or "is the person *pointing?*" would be precursors to starting to listen, if the former case were true, and, in case of the latter, looking in the direction of the pointing arm/hand/finger to find what is being pointed at. These examples constitute analysis of high-level *function.* Conversely, a listener's behavior of looking in the pointed direction is a sign to the speaker that s/he knows that the gesture is a deictic one, and has correctly extracted the relevant direction. In this example, the gaze behavior resulting from correct functional analysis can serve double duty as direct feedback, and constitutes therefore efficient process control.[6] *Functional analysis*—determining the function of a

1969], i.e. as the goal-directed use of communicative acts in context.

[5] Since turn changes between two speakers can happen with no gaps (0 ms pauses) [Goodwin 1981] we would need some sort of prediction mechanism to simulate this feature.

[6] It would also be correct and efficient feedback if an agent erroneously concluded that the gesture was iconic and therefore looked at the speakers hand instead, since this would clearly indicate to the speaker the error made. Interestingly, in this case the generation of correct feedback coincides with the actions necessary for further interpretation of the input.

ACTIVITY		
AUDITORY*	VISIBLE**	MULTIMODAL
Speaking	Gesturing	Paying attention
Assertive	Deictic	Addressing me
Directive *(commands)*	Iconic	Giving turn
Commissive	Pantomimic	Taking turn
Declarative	Symbolic	Wanting turn
Expressive	Butterworth♥	
Back channel	Self-adjustor	
Filler♠		

*See Searle [1969] for a treatment of speech acts. **This applies to both facial and manual gestures [Rimé & Schiaratura 1991, Ekman 1979, Effron 1941]. ♠Also referred to as "filled pause;" utterances like "aaaah" and "uuuuuh." ♥The gestural equivalent to filled pauses.

Table 1. Given a speaker's activity in the vocal tract, body, head, face or hands, a listener must classify this (find a behavior's function) to participate successfully in the dialogue. These are some high-level functions of multimodal actions that need to be recognized. It may be noted that most utterances directed to a computer agent would probably be directive.

multimodal action—is thus a *precursor* to both content analysis and correct feedback generation. Let's look at another example, using only the speech mode. The following exchange may look perfectly fine:

```
A: So, my funds will be withdrawn.     (1)
B: I'm so sorry to hear that!
```

until we add the accompanying intonation, which goes up at the end of the word "withdrawn" as indicated with a question mark:

```
A: So, my funds will be withdrawn?     (2)
B: I'm so sorry to hear that!
```

We find B's response inappropriate and would infer that B thought A was making a remark, not asking a question. If B had "computed" the correct function for A's utterance, (i.e. *question*) her response would probably have been different, along the lines of "No, they won't!" or "I don't know." Psychological research of the past decades has identified a number of the functions that need to be recognized, a selection of which is shown in Table 1.

The issue of functional analysis is neither one of computational power, nor of top-down/bottom-up processing; it is a sequential issue. Nothing prevents the use of either top-down or bottom up analysis to extract functional attributes of a speaker's behavior, and adding computational power will certainly speed up the process of analysis. But neither will eliminate the sequential dependency between the two steps of determining an action's function and analyzing its (possible) meaning(s). The second reason why this dependency is important is simple: More assumptions can be made with global information than local—an agent can do a lot more with general information when details are missing than with detailed information when the global perspective is lost.[7] By giving the highest-level functions highest priority, the most useful responses can be generated even if other information is missing, resulting in increased robustness.

The functional aspects of face-to-face interaction can, and should, be extracted by means of *multimodal* analysis; that is, any feature, body part, intonational cue or even lexical analysis could assist in the process. A major part of creating multimodal computer agents is finding how to extract the necessary information.

The Link Between Functional Analysis and Process Control

As we saw in the pointing example, correct and relevant feedback generation often follows automatically from correct functional analysis when the interactors are both human. For an artificial agent, however, we need to specifically address two issues that are given in the human-human case. The first is that we need to model the agent *in our own image*, i.e. with a head,

[7] This can be seen by a simple example: If I know that the speaker has just pointed in a direction and asked me a question, I can look in that direction and search for a likely referent about which the speaker may be asking. If I think the gesture is pantomimic, however, I will not look in the direction pointed, and am unlikely to find the referent as quickly, since I will have to correct my (functional analysis) error before being able to search for the referent. (See also interaction examples 1 and 2 in text).

face, gaze, arms, hands, and a body. This is because in face-to-face interaction, sensory organs, bodily constraints, attentional and mental limitations are linked together in a way that is intimately integrated and provides dialogue with an intricate feedback mechanism, the absence of which has been shown to disrupt discourse [Nespoulous & Lecours 1986].[8] In other words, if any parts of this mechanism are broken or missing, dialogue may break down.[9]

The second condition we need to fulfill is, as mentioned before, that the behaviors produced by the system be guaranteed execution within a given time limit, as determined by the pace of the dialogue. Because dialogue state is constantly changing, we need a mechanism that ensures that behaviors be executed at the time they are relevant—not before and not after.

YMIR:[10] An Architecture for Multimodal Agents

An architecture is being developed, called †mir, that directly addresses the above issues. Ymir is a hybrid system, based on the three-layer model of multimodal dialogue, and incorporates features of a black-board approach [Nii 1989, Selfridge 1959]: multiple knowledge sources cooperate to provide a solution to a problem—in this case to interpret user actions and generate appropriate responses. Psychosocial expertise (dialogue management) is separated from the main interpretive process (content interpreter) in a modular fashion [Bolt, personal communication, 1992, Walker & Whittaker 1990]. The model also borrows some insights from behavior-based AI [Mayer & Wilson 1991], particularly those expressed by Maes [1990, 1989].

Psychosocial skills in this system are divided into two subsystems, a *Reactive Layer* and a *Process Control Layer* (PCL). These systems can issue high-level action requests based on the dialogue state, interpretation and input variables. Motor actions are computed from these high-level requests by an *Action Scheduler* (AS), using a knowledge base of motor schemes and motor combination rules.

In accordance with the requirement for real-time performance, the behaviors that are most time-specific take precedence in terms of input analysis *and* execution. These are handled by the Reactive Layer:

1. Signals related to attentional focus (gaze being the primary indicator).
2. Back-channel feedback (symbolic head motions and speech).
3. Signals related to turn-taking (mostly gaze and facial/head gestures).

The PCL is concerned with delivering higher-level output related to the dialogue and internal process regulation, such as asking questions when problems arise and regulating the interpretation modules. It is also involved in maintaining the dialogue history. The PCL receives reports from a topic interpreter about the status of the interpretation of recent user behaviors and issues action requests that depend on the results of that interpretation.

DISCUSSION

A particularly interesting feature of this architecture is the way reactive responses can be modeled separately from reflective actions. This simplifies construction of behaviors and allows for their incremental development. Another important feature is how behaviors are separated from their exact morphology, which can instead be computed at runtime, taking time constraints and current state of motors into consideration. Finally, the separation of dialogue management and topic knowledge/interpretation has the potential to accommodate multiple knowledge

[8] Nespoulous & Lecours [1986, page 61] say: "... Dahan [see ref., op. cit.] convincingly demonstrated that the absence of regulatory gestures in the behavior of the listener could lead the speaker to interrupt his speech or to produce incoherent discourse."

[9] It is also possible that some violations can be fixed with clever engineering of the agent behavior, its visual appearence or its environment.

[10] According to Nordic religion, as told in Icelandic sagas [Sturluson 1300~1325], †mir (pronounced e-mir, with the accent on the first syllable) was a giant who was killed by the gods Ó›inn, Vili and Vé, who subsequently used his carcass to make heaven and earth. The earth then became a source of many new imaginative humanoid lifeforms.

bases without changing the structure of the underlying dialogue mechanisms. The system is currently being used to design prototype interface agents and preliminary results are promising.

SUMMARY

This paper outlined characteristics of multimodal interaction important to computational modeling of autonomous characters and provided arguments for an architecture that combines reactive and reflective behaviors, where functional analysis is a precursor to content analysis and feedback generation. A testbed architecture for psychosocial dialogue skills, Ymir, was shortly described. This architecture's main features are a layered approach containing separate modules for reactive behaviors, process control behaviors, topic-related interpretation and action execution.

Acknowledgments

This paper has benefited greatly from interactions and discussions with Richard Bolt, Marilyn Walker, Justine Cassell, Steve Whittaker, Pattie Maes and Joshua Bers, and for this they have my sincere gratitude. Justine Cassell, Steve Whittaker and Pattie Maes were generous enough to offer valuable comments on drafts of this paper. Joshua Bers has also my thanks for always being willing to discuss anything from the general to the particular. This work was sponsored by the Advanced Research Projects Agency (ARPA) under Rome Laboratory contract no. F30602-92-C-0141, and by Thomson-CSF.

REFERENCES

Argyle, M. & Cook, M. (1976). *Gaze and Mutual Gaze*. England: Cambridge University Press.

Argyle, M., Lefebvre, L., & Cook, M. (1974). The meaning of five patterns of Gaze. *European Journal of Social Psychology, 4(2)*, 125-136.

Bolt, R. A. (1985). Conversing With Computers. *Technology Review*, Feb./March. Cambridge, MA: MIT Press.

Bolt, R. A. (1987). The Integrated Multimodal Interface. *The Transactions of the Institute of Electronics, Information and Communication Engineers (Japan), Nov., J79-D(11)*, 2017-2025.

Effron, D. (1941/1972). *Gesture, Race and Culture*. The Hague: Mouton.

Ekman, P. (1979). About Brows: Emotional and Conversational Signals. In M. von Cranach, K. Foppa, W. Lepenies, & D. Ploog (eds.), *Human Ethology*, 169-249.

Ekman, P. & Friesen, W. (1969). The Repertoire of Non-Verbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica, 1*, 49-98.

Goodwin, C. (1986). Gestures as a Resource for the Organization of Mutual Orientation. *Semiotica*, 62(1/2), 29-49.

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Kahneman, D. (1973). *Attention and Effort*. New Jersey: Prentice-Hall, Inc.□

Kleinke, C. (1986). Gaze and Eye Contact: A Research Review. *Psychological Bulletin, 100(1)*, 78-100.□

Koons, D. B., Sparrell, C. J. & Thórisson, K. R. (1993). Integrating Simultaneous Input from Speech, Gaze and Hand Gestures. Chapter 11 in M. T. Maybury (ed.), *Intelligent Multi-Media Interfaces*, 252-276. Cambridge, MA: AAAI Press/M.I.T. Press.

Laurel, B. (1990). Interface agents: Metaphors with character. In B. Laurel (ed.) *The Art of Human-Computer Interface Design,* 355-365. Reading, MA: Addison-Wesley Publishing Co.

Lord, C. & Haith, M. M. (1974). The Perception of Eye Contact. *Perception & Psychophysics, 16(3)*, 413-16.

Maes, P. (1989). How to Do the Right Thing. A.I. Memo No. 1180, December, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Maes, P. (ed.) (1990). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. Cambridge, MA: MIT Press/Elsevier.

Mayer, J. A. & Wilson, S. (eds.) (1991). *From Animals to Animats*. Cambridge, MA: MIT Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago, IL: University of Chicago Press.

Nespolous, J-L & Lecours, A. R. (1986). Gestures: Nature and Function. In J-L Nespolous, P. Perron & A. R. Lecours (eds.), *The Biological Foundations of Gestures: Motor and Semiotic Aspects,* 49-62. Hillsdale, NJ: Lawrence Earlbaum Associates.

Nii, P. (1989). Blackboard Systems. In A. Barr, P. R. Cohen & E. A. Feigenbaum (eds.), *The Handbook of Artificial Intelligence*, Vol. IV, 1-74. Reading, MA: Addison-Wesley Publishing Co.

Pierrehumbert, J. & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (eds.), *Intentions in Communication.* Cambridge: MIT Press.

Rimé, B. & Schiaratura, L. (1991). Gesture and Speech. In R. S. Feldman & B. Rimé, *Fundamentals of Nonverbal Behavior*, 239-281. New York: Press Syndicate of the University of Cambridge.□

Selfridge, O. G. (1959). Pandemonium: A Paradigm for learning. *Proceedings of the Symposium on the Mechanization of Thought Processes*, 511-529.

Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. London: Cambridge Univ. Press.

Sturluson, S. (1300~1325). *Edda.* Á. Björnsson prepared for publication. Reykjavík: I›unn, 1975.

Thórisson, K. R. (1993). Dialogue Control in Social Interface Agents. *InterCHI Adjunct Proceedings '93*, Amsterdam, April, 139-140.

Thórisson, K. R. (1994). Face-to-Face Communication with Computer Agents. *AAAI Spring Symposium on Believable Agents Working Notes*, Stanford University, California, March 19-20, 86-90.

Walker, M. & Whittaker, S. (1990). Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. *29th Annual Proceedings of the Association of Computational Linguistics*, 70-78.

Yngve, V. H. (1970). On Getting a Word in Edgewise. *Papers from the Sixth Regional Meeting*. Chicago Linguistics Society, 567-78.