Introduction

1.

We are creating a new arena of human action: communication *with* machines rather than operation *of* machines.

—Card, Moran & Newell (1983, p. 7)

As humans, we are naturally endowed with multimodal input/output capabilities. Multimodal interactions happen between people most every day: we exchange glances, gesture to each other, speak and make facial expressions. The purpose of these interactions is usually to communicate certain information to, and receive information from others. As any student of psychology will know, multimodal I/O as it happens in face-to-face interaction is a complex phenomenon and many of its features and smaller pieces make valid research topics and research fields. Yet most people, when asked about how they manage to communicate complex information in a short face-to-face interaction, they shrug and reply "It's easy—getting a machine to do that should be trivial" (or even worse "Haven't they done that already?"). In a paper on computers and common sense, Phil Agre [1985, p. 72] writes:

> Playing chess is easy, but making breakfast is enormously complicated. This complexity stares us in the face every morning, yet it is invisible.

Face-to-face interaction is like making breakfast. It looks easy. But when it comes to making a computer do the same, things start getting mighty complicated.

Here, the approach taken to this problem is not in the typical tradition of divide-and-conquer, but instead to look at multimodal interaction holistically, with the purpose of constructing a computer system that can sustain and support such interactions with a human. To this end I have designed an architecture that allows for the construction of multimodal agents—agents that can interact with people using speech, gesture and gaze. I have also built a prototype agent in this architecture. These will be discussed in Chapters 7., 8. and 9. In this chapter we will define some important terms, take a close look at the goals of this work and give an overview of the rest of the thesis.

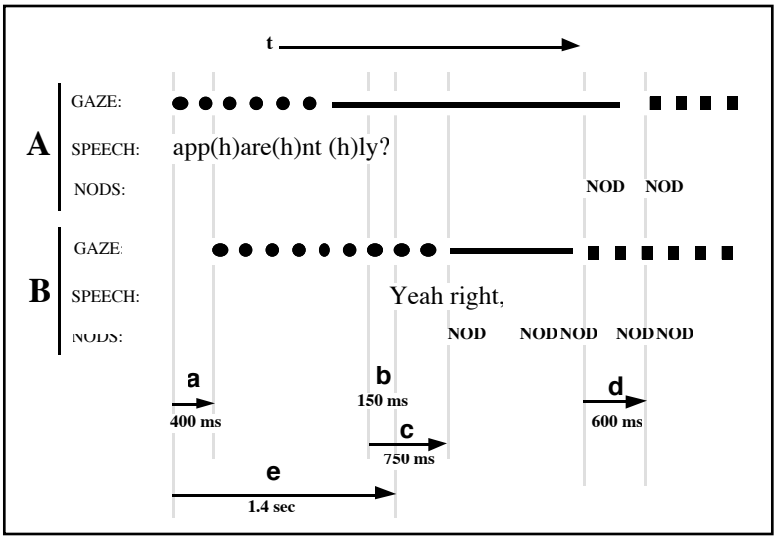

FIGURE 1-1. Transcript spanning 3 seconds of a typical two-person conversation, showing the timing of speech, gaze and head nods for each conversant (Adapted from Goodwin [1981]). "A brings her gaze to the recipient. B reacts to this by immediately bringing her own gaze to A. The two nod together and 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.

Circles indicate gaze moving toward other, lines indicate a fixation on other, squares are withdrawal of gaze form other, question mark shows rising intonation.

1.1 What is Needed

The transcription in Figure 1-1 demonstrates the complex nature of face-to-face discourse [Goodwin 1981]. Here, rapid responses and more reflective ones are interwoven in a complex pattern. Person A and person B exchange glances that are timed to the decisecond; they give each other feedback and take turns speaking with admirable efficiency. People are obviously very good at doing this, and to date no computer system has been able to replace one of the participants and produce the same pattern as shown in the example. This is because a system that can do this needs to be responsive to the environment, yet be capable of longer-term planning. Moreover, it needs to keep track of multiple sources of information including a person's gaze, facial expression, gesture, intonation, body language, in addition to speech content.

To date, research has fallen short when it comes to these essential topics in face-to-face interaction:

1. *Continuous-input over multiple modes.*
2. *Integration of multimodal inputs.*
3. *Coordination of actions at multiple levels of granularity.*
4. *Bridging between sensory input and action output.*

Instead of a "vending-machine" interaction style (communicate all information ... wait for system response), continuous input allows a system to support interruptions, incremental input and incremental interpretation. Multimodal input contains multiple data types; these have to be integrated in some manner to support correct feedback generation. In dialogue, real-time responses are tightly coupled with more "reflective" ones; "um"s and "ahh"s are automatically inserted while we think of what to say. How we allow a machine to do this as output is also an open question. A complete bridging between sensory input and motor output is necessary if we want to have a platform that allows us to experiment with various designs for humanoid agents.

1.2 Goals of This Work

"Designing computers that are to operate in isolation is one thing, but designing computers that are to occupy an important place in the lives of real people is something else."

—Philip Agre (1994, p. 230)

This thesis describes the efforts of endowing a multimodal, on-screen computer agent with psychosocial dialogue skills aimed at supporting and sustaining dialogue with a human. Two closely related problems or issues are addressed by this work. The first is the general issue of human-computer interaction. The new type of interface proposed takes advantage of people's knowledge about face-to-face interaction, turn-taking and perceptual abilities of interacting parties to provide a consistent metaphor for the relationship between human and computer. By introducing a situated social entity into the human-computer relationship, enabling full-duplex multimodal interaction, a number of benefits may be expected, among them increased flexibility and greater reliability in the interaction sequence. The resulting agent-based system will provide a powerful and intuitive new means for interacting with computers and have potential application in a multitude of systems requiring high-level command.

The second issue addressed is that of dialogue modeling. In order for the multimodal interface agent metaphor to work, the agent has to be capable of a minimum set of skills: its underlying mechanism has to capture elements that are critical to the structure of multimodal dialogue, such as gestural signals, body language, turn-taking, etc., and integrate these in a way that works. I propose a computational architecture of psychosocial dialogue skills, called Ymir, that bridges between

multimodal input analysis and multimodal output generation. A character has been built in this architecture, called Gandalf, that can interact with humans in real-time, perceiving and generating various mutlimodal actions. By testing this character experimentally with human subjects, the validity of the approach is evaluated in various aspects.

1.2.1 Terms & Definitions

A few words on important terms are in order, without diving into the bottomless pit of definitions. The following terms are in special need of discussion: "Multimodal," "interface," "agent", "humanoid" and "psychosocial skills". The term "mode" as used here generally refers to an anatomically separate mechanism on the human body, or mechanisms carrying different kinds of data, enlisted for the purpose of communication with other humans, such as gesture and speech; intonation and body language, etc. "Multimodal" means therefore the collection of many such mechanisms. "Interface" traditionally means the place where two different systems meet: here it is the human and machine that meet, hence the term "human-machine interface." The term "agent" has served numerous meanings, but can be considered here to mean broadly "the conceptual categorization of one or more computer-stored goals, and the collective capability to carry out those goals, to the computer user's interest." A vacuum-cleaning robot would be a good example of an agent according to this definition. As we will see later, this is a slightly too broad definition for the current purposes, but it will do for now. A "humanoid" is that which duplicates many human characteristics, yet is *not* human. The distinction that is being emphasized by using this word is the one between animals, insects and related creatures on the one hand, and human-like creatures on the other. To be grouped with the latter one would have to share with humans at least some of our unique features: a human face, language understanding and generation, social skills, among other things. Lastly, "psychosocial skills" are the skills needed to orchestrate, co-operatively, goal-driven communicative interaction with other agents. The current work is thus a contribution to the broad scope of dialogue management, rather than narrower aspects or smaller parts of dialogue such as language understanding, gesture recognition or agent animation.

Since the emphasis here is on the full loop of multimodal input analysis and multimodal output generation, a number of assumptions have been made and gaps filled where research was lacking or too unwieldy for a one-man project. These include knowledge representation, linguistic issues, cognitive modeling and philosophical questions of all sorts. I hope the reader can forgive these unavoidable gaps in my treatise, and ask that you try to focus on the problem of full-duplex interaction, which, in my opinion, should be the starting point for all other issues of

dialogue. We can then leave it to future research to fill in the missing details.

1.2.2 Outline of Thesis

The first 3 chapters present background material: Chapter 2. discusses the face-to-face metaphor, Chapter 3. reviews the psychological research in multimodal communication and multimodal computer interfaces and Chapter 4. gives an overview of related research on software agents, robots and artificial intelligence.

> The skills themselves are basic: breaking eye contact when you want to speak; noting whether the other person is looking in the right spot when you point something out to them; describing things and events with your hands... Can such general, practical conversational expertise be imparted to computers?
>
> —Richard A. Bolt (1987, p. 2024)

Chapters 5., 6. and 7. present the approach taken here to creating interactive, humanoid characters, and the underlying assumptions. Chapter 5., "Computational Characteristics of Psychosocial Dialogue Skills", focuses on the hard issues in multimodal dialogue, their computational characteristics and ways to formalize these for implementation. A three-layer feedback model of multimodal dialogue is introduced that addresses its real-time constraints and mode integration. Chapter 6., "J.Jr.: A Study in Reactivity", describes a pilot system that served to explore the issues of real-time dialogue feedback, back channel and turn taking. The limitations that emerge from this study motivate many features of Ymir[1]—presented in Chapters 7. and 8.—a generative model for a communicative agent's sensory, decision and motor processes. Chapter 9. describes the first character created in Ymir, Gandalf. Chapter 10. presents the results of an evaluation of Ymir/Gandalf using human subjects, and discusses the methods used to create humanoid agents in Ymir.

Chapter 11. discusses validity in the design of multimodal agent-based interfaces and relates these to possible implementations of communicative humanoids. General conclusions from this work are drawn, and directions for future work given, in Chapter 12.

1. Pronounced "***e***-mir" with the accent on the first syllable. The name comes from Nordic religion; see side bar page 89 for background.