Computational Simulations of Mediated Face-to-Face Multimodal Communication

by

Melanie A. Baljko

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Your file Votre référence
ISBN: 0-612-94270-8
Our file Notre référence
ISBN: 0-612-94270-8

Abstract

Computational Simulations of Mediated Face-to-Face Multimodal Communication

Melanie A. Baljko
Doctor of Philosophy, November 2004
Graduate Department of Computer Science, University of Toronto

Individuals who have little or no functional speech due to underlying physical disorder may instead use a computational device, called a Voice Output Communication Aid (VOCA), to produce synthesized speech. We describe a previously-unidentified set of tradeoffs that face the designers of Augmentative and Alternative Communication (AAC) systems, of which VOCAs are one possible component. On the one hand, the mode of synthesized speech that is afforded by a VOCA can be used to produce communicative actions that are more likely to be successfully interpreted than those produced using other modes, especially by unfamiliar communication partners — a benefit that can justify the often-sizeable effort that must be expended by individuals in order to use their VOCAs. On the other hand, the use of this so-called *aided* mode can conflict with the simultaneous use of the other, *unaided* modes, such as facial expression, eye gaze, vocalization, and gesture — a negative effect on the interlocutor's ability to produce *multimodal* communicative actions. These actions can be equally or even more effective than unimodal ones produced using synthesized speech alone, while also requiring less effort.

The use of multimodal interfaces for VOCAs was first proposed by Shein et al. [1990]; prototypes have since been developed by Treviranus et al. [1991], Smith et al. [1996], and Keates and Robinson [1998]. We describe and formalize a previously-unidentified mechanism whereby a repertoire of modes of articulation affords a repertoire of mode strategies. The mechanism developed here accounts for the effects of *conflict* among the modes in a repertoire — a situation in which two or more modes rely on common underlying communicative effectors, as is the case with synthesized speech and gesture. We instantiated the mechanism computationally and used it to simulate the consequences of unimodal and multimodal VOCA interfaces on a simulated communicator's repertoire of mode strategies. We show that, for the unimodal interfaces, empirical and anecdotal evidence agree with the simulation results. We also show, through the simulations, that the mechanism of mode conflict can have serious consequences for the utility of multimodal VOCA interfaces and thus the bottleneck-reduction hypothesis.

Acknowledgments

I would like to express my sincere thanks to Professors Graeme Hirst and Fraser Shein, who, from the outset, agreed to a co-supervisory arrangement. This arrangement entailed considerably more time and effort on their parts than the conventional single-supervisor arrangement, but I sincerely believe that this thread of inquiry would not have been possible otherwise. I thank each for their time, effort, support, patience, and commitment. Financial support came from both co-supervisors, through their grants from the Natural Sciences and Engineering Research Council of Canada, and the grant "Advanced Computer Access for People with Disabilities" in particular.

I would like to express my sincere thanks to Professors Ronald Baecker and Mark Chignell, who were active members of my thesis advisory committee and whose feedback was very beneficial. I am also very grateful to Suzanne Stevenson, who provided very thorough and constructive feedback throughout the many phases of this work and who was an especially thorough reader. I am very fortunate that she agreed to be a member of my advisory committee and am thankful for her advice and counsel on the non-dissertation related aspects of graduate school and academic life.

I am very thankful to Professor Norman Alm, who made the trip from University of Dundee. As external examiner, he was thorough and constructive in his comments. I am also thankful to Professor Tom Chau from the Institute of Biomaterials and Biomechanical Engineering, who also was a thorough and careful examiner.

For their support, I am thankful to my parents and to my brother and sister (my sister Christine deserves special thanks for reading Chapter 7 while on vacation in Kukljica, which went well beyond the call of family duty). I am dedicating this dissertation to my maternal grandparents, Irma Wydooghe and Joseph Gheysens, and to my paternal grandparents, Rajka Lonić and Anđelo Baljko, who made choices that created opportunities and effects that eventually made it possible for me to enjoy this big luxury — the chance to pursue an education out of interest and curiosity. I am thankful to all of my office-mates throughout the years, whose company made graduate school very memorable indeed: Daniel Marcu, Cathy Jansen, and Natasa Przulj, and I am very grateful to all of my other wonderful friends who didn't actually have to share office space with me. And the best for last — Vassilios "JYP" Tzerpos accompanied me throughout all of this rather long journey; one couldn't haved asked for a better partner.

Contents

Chapter 1

Introduction

1.1 Motivation and Research Issues

Individuals who have little or no functional speech due to underlying physical disorder may instead use a computational device to produce synthesized speech. Such devices, which are typically referred to as Voice Output Communication Aids (VOCAs), are but one type of Augmentative and Alternative Communication (AAC) device. Furthermore, AAC devices are but one component of AAC systems; AAC systems also have, among other components, a repertoire of communication strategies, which is typically adapted to each individual and his or her communication partners.

Using a VOCA, an individual can "compose" an intermediate representation of a targeted spoken utterance by selecting a sequence of symbols (e.g., words that are spelled out by sequences of letters, or phrases that are given by sequences of words or sequences of icons). Typically, VOCAs are implemented with pre-stored vocabularies, from which words and phrases can be selected. The intermediate representation can then be used either as the input to a text-to-speech module or as a retreival key to an inventory of pre-stored samples of digitized speech. A VOCA is shown in figure 1.1 below.

Initial and ongoing research efforts have focused on improvements to VOCAs and have yielded incremental improvements with respect to AAC system effectiveness. However, improvements are yet needed in order to achieve desired levels of AAC system effectiveness (e.g., to mitigate more completely the functional limitations that individuals experience with respect to their ability to engage in communicative exchanges).

In the last decade or so, some researchers have shifted their focus to the creation of VOCAs with *multimodal* interfaces. This shift in focus has been spurred by the recognition that one of the major problems with using the mode of synthesized speech is *production latency*. In order to produce an utterance using synthesized speech, the user must build a representation of the intended utterance, which most often is represented textually and is subsequently passed to the VOCA's text-to-speech module. The complexity of the utterances that can be articulated using synthesized speech is dependent on the types of textual representations that can be constructed, which, in turn, are most often determined by the complexity of the pre-stored vocabulary and the user's ability to navigate and select elements from it. (Alternatively, the user may enter words directly, often with

the assistance of word-prediction software.)

Navigation through and selection from the pre-stored vocabulary is accomplished through a series of input actions such as successive key presses or successive eyebrow raises. Many different types of volitional movement, through the use of an appropriate input device, can be harnessed as sources of input actions. These actions, however, require time and effort to produce and thus a period of latency results, during which the utterance is composed before it is articulated by the device. This production latency can be burdensome for the communication partner and can be easily misunderstood (and hence be disruptive to the interlocutor's ability to take and to keep the conversational turn). The problem of production latency could be ameliorated if the user could provide, over a given period of time, more information to the device (where information is typically given by number and type of selectional input actions). Since the information is provided to the device via its interface, this desire is typically expressed as a need for the interface to have a greater "information bandwidth." If the user could provide multiple input actions (possibly even simultaneously), then this bandwidth could be increased. The *bottleneck reduction hypothesis* [Shein et al., 1990] is that AAC devices with multimodal interfaces will offer advantages over unimodal ones to individuals with communication disorders (so-called *aided communicators*). Research efforts are underway to develop such devices (e.g., Treviranus et al. [1991]; Roy [1992]; Roy et al. [1993a,b, 1994c,b]; Smith et al. [1996]).

1.2 Goals of this Research

The long-term goal of this research is to develop more effective computational interventions for communication disorders. Two goals are more immediate: to analyze and to evaluate the bottleneck reduction hypothesis, and, on the basis of that evaluation, to identify and formalize the factors that determine the effectiveness of multimodal interfaces for computational AAC devices.

Figure 1.1 A Voice-Output Communication Aid (VOCA) in use. Image used with permission, in accordance with the AT/AAC enABLES Copyright & Use Policy of the Department of Speech & Hearing Sciences at the University of Washington (2004).

1.2.1 Analysis of the Bottleneck Reduction Hypothesis

The motivation for the development of multimodal interfaces is based on the predictions of the bottleneck reduction hypothesis. We will examine the hypothesis' rationale and identify the theoretical model of communication upon which it is based. This examination will show that the rationale is based on an information-theoretic model of communication and, moreover, that it makes the faulty assumption that the communicative process is wholly mediated by the AAC device. We will show, through the application of Clark's characterization of the process of communication [1996], that it is the *AAC system* that truly mediates the communicative process and that the AAC device only partially mediates the performance of some communicative actions by interlocutors. Given this, we need an alternative characterization of the role of the AAC device. I will describe the various characterizations of VOCAs in the research literature and show that the most valid characterization is one in which the VOCA is seen as serving *to augment* an interlocutor's repertoire of modes. The repertoire of modes, in turn, *affords* a repertoire of mode strategies.

1.2.2 An Explanatory Model of the Production of Multimodal Communicative Action

The correspondence between an individual's repertoire of modes and his or her repertoire of strategies is highly relevant for the design of AAC systems; the design rationales for clinical interventions are based, in part, on exploiting this correspondence (i.e., they augment an individual's repertoire of modes in order to augment his or her repertoire of strategies).

I will develop a mechanism which establishes the connection between an interlocutor's repertoire of modes and the strategies of mode use in the production of multimodal communicative actions. I will focus on the mode strategies employed by individuals who use VOCAs, and I will analyse the consequences of a multimodal VOCA interface on the individual's mode strategies. In order to do so, it is necessary to define what is meant by a *mode strategy*. But even before this, we must define what is meant by a *mode*. We will examine how the notion of a *mode of communication* has been defined in the past (Chapter 2), and then define it for our research. The definition that will be developed here is derived from first principles, and I show that it is not appropriate for characterizing an individual's communicative actions: a mode of communication, as it turns out, can only be used by participants in joint action.

When engaged in a communicative exchange, interlocutors can be seen as making use of the available resources (their bodies and any artefacts available in the environment) to produce behaviours that are observable by others. These behaviours are often discrete and have been variously labeled as *speech acts* [Austin, 1962; Searle, 1969], *conversation acts* [Traum and Hinkelman, 1992], *dialogue acts* [Bunt et al., 1995], *conversational moves* [Power, 1979], and *signals* [Clark, 1996]. However these actions are labeled, they form the *interposed elements* that mediate between individuals engaged in a process of communication [Craig, 1999]. We characterize these interposed elements as a set of temporally-coordinated sub-actions that are each specific to a particular mode (e.g., such as sub-actions articulated using speech, eye gaze, facial expression, and gesture of the hands, torso, and head).

The interposed elements of communicative processes are discrete, multimodal communicative

actions. An interlocutor's own body provides the resources which make the modes available (e.g., the *body effectors* that underlie the use of speech, gesture, facial expression, and so on). Interlocutors who do not have physical disorders have unfettered use of their body effectors (although effects due to fatigue do occur). Interlocutors who have physical disorders can be seen as experiencing constraints with respect to their body effectors, which in turn affects the availability of their modes. I will develop a framework in which VOCAs are characterized as affording an additional *aided* mode of synthesized speech. Empirical evidence shows that aided communicators make use of *all* of their modes of communication — both the aided mode and the unaided modes. Furthermore, these modes are not necessarily used in isolation; in fact, multiple modes are typically used, combined in a variety of ways. It is more accurate to characterize individuals as using *mode strategies*, rather than making use of modes, per se.[1] A mode strategy is not just which particular combination of modes is used, but also how the modes are used (e.g., whether *redundancy* has been used or not, a notion that will be defined in subsequent chapters). The selection of mode strategy is an issue not only for individuals who use communication devices (and who have both aided and unaided modes), but for interlocutors in general. There is a multitude of different mode strategies; their identification and classification will be discussed in Chapter 2 and is an area of ongoing research.

An individual with a communication disorder, even if not using an AAC device, has an already-existing repertoire of modes, which, in turn, affords a *repertoire of mode strategies*. This repertoire consists of *unaided-unimodal* strategies (i.e., the use of an unaided mode in isolation) and *unaided-multimodal* strategies (i.e., the use of unaided modes in combination). The use of an AAC device — whether it has a unimodal or multimodal interface — *affords* the additional *aided* mode of synthesized speech. Thus, an individual who uses such a device *augments* his or her repertoire of mode strategies with the *aided-unimodal* mode strategy (i.e., the use of the aided mode in isolation) and various *aided-multimodal* strategies (i.e., the use of the aided mode in various combinations with the other unaided modes).

1.2.3 An Analysis of the Design Process for AAC Systems

The stated goal of an AAC intervention is to circumvent the limitations that an individual experiences with respect to his or her ability to fulfill social roles or to perform tasks. I will examine how this intervention goal is accomplished and will show that approach is to consider the single, solitary *unimodal-aided* mode strategy as a replacement for the individual's already-existing repertoire, and then to focus on improving and further developing this mode strategy. However, I will argue that a better way to accomplish the intervention goal is to augment the individual's repertoire of mode strategies (e.g., so that the repertoire is serviceable in all of the scenarios and with all of the communication partners that the individual will encounter). Furthermore, I will analyse the design process for AAC systems, in order to determine how this goal is met and how it might incorporate the design of multimodal VOCAs in the future.

I will show that there are subtle issues in the augmentation of mode strategy repertoire. In particular, the operation of a typical, unimodal VOCA requires the individual to look down, which means that the mode of gaze cannot be used simultaneously. Also, the AAC device requires input

[1] E.g., the use of the mode of synthesized speech entails the use of a mode strategy (namely, the isolated use of the mode).

actions (typically performed by the hands), which means that gesture cannot be used simultaneously. Thus, the use of the aided mode *conflicts* with the simultaneous use of the other modes. This means that if a multimodal-aided strategy is to be used, the use of the unaided modes must either precede or follow the use of the composition phase required for the aided mode. Typically, the use of the unaided modes follows the composition phase (for instance, the unaided modes are used to modulate the meaning of the synthesized speech). According to the bottleneck-reduction hypothesis, a multimodal device is expected to afford a more effective unimodal-aided mode strategy. However, I will show that what the hypothesis does not acknowledge is that the use of the multimodal device will conflict more with the other modes, and thus, reduce (relative to the unimodal device) the interlocutor's repertoire of multimodal-aided mode strategies. I will show that this *global* assessment of an AAC device's impact (i.e., an evaluation in terms of the device's impact on an individual's repertoire of mode strategies) is crucial to the evaluation of an AAC system's effectiveness (as opposed to its *local* effectiveness — its utility in a particular communicative context). I will also show that a computational simulation tool, provided appropriate techniques to characterize mode strategies are developed, can be used to evaluate global effectiveness, and thus can complement existing, empirical evaluation techniques. The need for such simulation tools is a logical consequence of the increasingly apparent need for a more principled approach to AAC system design [Light, 1999].

1.2.4 A Technique for Characterizating Strategies of Mode Use

As described above, a VOCA serves to augment an individual's repertoire of strategies of mode use (albeit indirectly — the VOCA affords an additional mode of articulation, and the augmented repertoire of modes affords an augmented repertoire of mode strategies). I will show that, given a particular communicative context, each mode strategy has its advantages and its disadvantages. For instance, the unimodal-aided mode strategy (i.e., the use of synthesized speech in isolation) has an important advantage over the other mode strategies, such as those that are based on gaze, facial expression, vocalizations, and gesture: using it, a individual can accomplish things that would otherwise be difficult or not possible at all through the use of the other mode strategies. But this mode strategy can be physically demanding and fatiguing to use. Furthermore, this mode strategy also might not be necessary with certain communication partners (e.g., the other mode strategies might suffice).

I will identify different consequence attributes[2] (e.g., with respect to fatigue and to understandability) that follow from the use of a mode strategy and describe a technique for combining the attribute-specific values into a single value, which, in turn, can be used to characterize the mode strategy. By applying the characterization technique to each mode strategy in an interlocutor's repertoire, I will characterize the interlocutor's repertoire in a way that permits comparison with other possible repertoires. I will show that one way to compare and to contrast various AAC devices (such as unimodal and multimodal AAC devices) is with respect to their relative impacts on

[2]In the literature on decision theory, and in later parts of this dissertation, a course of action (e.g., such as the use of a particular mode strategy) is described as having *one* consequence which has multiple attributes (as opposed to a course of action having several consequences). Furthermore, each of the attributes of the consequence has *value*, which is used to quantify benefit or shortcoming.

an individual's repertoire of mode strategies.

We will introduce the notion of mode strategy selection: an aided communicator, when designing his or her communicative actions, is considered to be a decision maker who chooses among a set of alternatives in light of their possible consequences. I will address the question of which alternatives (i.e., strategies of mode use) are afforded to the aided communicator, in the context of a particular instance of decision making, as a function of his or her repertoire of modes. I will also address the question of how the three models of decision making (decision under certainty, decision under risk, decision under uncertainty) might be applied to the task of choosing from among this set. Last, I will address the question of how should AAC devices be designed so that an individual's augmented repertoire of modes will afford "good" sets of alternatives in all of the contexts in which communicative actions are to be produced (i.e., in all of the contexts in which decisions are to be made).

1.2.5 Demonstration of *Bottleneck Reduction* by Computational Simulation

In section 1.2.3, the need for a computational simulation tool was identified, and in section 1.2.4, the need for techniques to characterize mode strategies was identified. As proof of concept, I will create a computational instantiation and demonstrate its utility in a novel application — a computational simulation tool that demonstrates the hypothesized relative merits of the various strategies of mode use that are available to a given individual (where that individual makes use of a given VOCA). I will represent a variety of degrees of physical disorder and a variety of unimodal and multimodal AAC devices. I will show that by manipulating the characteristics of the device (each corresponding to a possible AAC device design), the consequences of various possible designs can be demonstrated through computational simulation. I will also describe an evaluation technique that can be used to establish the usefulness of the computational tool, and apply it to the extent that is possible, given the currently available empirical data.

1.3 Overview of Contributions

The next five sections briefly outline the five main contributions of this dissertation.

1.3.1 Theoretical Foundation for Analyses of Multimodal Communicative Actions

The first major contribution of this dissertation is a new, theoretically-grounded definition of the modes of communication, and an analysis which distinguishes the notion of a mode of communication from the related, yet distinct, notions of a mode of articulation and a channel of communication. Previously in the research literature, the notion of the modes of communication has either been poorly defined or not defined at all.

The notion of a mode of communication is analysed and a definition is derived that is based on a formal model of the process of communication. According to this definition, a mode of communication can only be used by participants, collectively, in joint action. This notion of a mode of

communication is shown to be incompatible with the notion that is needed in order to characterize an individual's communicative actions as compositional (composites of multiple, mode-specific actions). An individual, ultimately, produces a communicative action autonomously; thus, the modes of communication cannot be used to characterize this autonomous action. To meet this need, the notion of a mode of articulation is defined and distinguished from a mode of communication; these are both clearly distinguished from the channels of communication. A criterion for the identification of multimodal synergy is given.

The dissertation proposes a novel model of the communicative resources that an interlocutor has for the production of multimodal utterances, and links this resource model to a formal model of the production process that is based on the mechanism of constraint satisfaction. Through this linking, the resources available to an interlocutor are set up as a *parameter* of the production process. This parameterization is a new feature, a feature that builds upon present formal models of multimodal utterance production. This process of constrained multimodal utterance production has not previously been modelled formally. In addition, previous formal models of multimodal utterance production have not accounted for the role of physical disorders.

1.3.2 An Alternative Model of AAC Interventions

The dissertation analyzes the characterization of dysfunction in communicative processes and shows that any such characterization necessarily entails the assumption of an underlying model of the process of communication. Three different characterizations of dysfunction are identified, each one based on a different model of the process of communication. One such characterization is based on a novel application of Clark's [1996] Contribution Model. This characterization of dysfunction suggests a different strategy for intervention, one based on augmenting the interlocutors' ability to collaborate, as opposed to attempting to augment directly their ability to communicate. This characterization contrasts with the information-theoretic characterization of dysfunction (i.e., as problems with information "through-put"), upon which the so-called bottleneck reduction strategy for AAC design is based. The dissertation provides an analysis of the merits of the bottleneck reduction strategy, and identifies the issue of *mode conflict*, which can have serious consequences for the conceptualization of the effectiveness of AAC systems.

According to the framework that will be established in chapter 3, the goal of clinical interventions for communication disorders is to augment an individual's repertoire of strategies, so that he or she is better able to achieve his or her goals (where the satisfaction of most of these goals, if not all of them, involves participation in communicative exchanges, or conversations[3] more generally). One approach is to augment the individual's repertoire of modes with an additional, *aided* mode.[4] As the next section will show, this novel characteristization of clinical intervention was foundational to the identification of a role for computer-assisted AAC design.

[3]Conversation is characterized as "the free exchange of turns among two or more participants" [Clark, 1996, p. 4]. The connection between the process of communication and the activities that individuals undertake to satisfy their goal is described in section 2.3.

[4]Other approaches include the removal or reduction of barriers to communication, adaptations to an individual's environment, and further development of an individual's natural abilities through training [Beukelman and Mirenda, 1998]. Clinical interventions are described in section 3.2.

1.3.3 Analysis of the AAC Design Process

The dissertation describes a novel analysis of the process whereby AAC systems are designed. In this analysis, the main phases of software design were identified in the AAC design process: requirements analysis, specification, implementation, and evaluation. This analysis yielded a much more concise, formal description of the inputs and parameters of the design process than has been previously described in the research literature. This more-specific formalization has provided the basis for identifying and formulating two different types of AAC system effectiveness, which have been labelled here *local* and *global*. Global effectiveness concerns the AAC system's impact on an individual's repertoire of mode strategies, for use across all relevant communicative scenarios. (Local effectiveness concerns the AAC system's impact on the individual's repertoire of modes strategies in a particular communicative scenario.) The importance of evaluating global effectiveness, as well as the difficulty in doing so, is the main motivation for developing computational simulation tools for use in AAC design. In effect, the dissertation identifies the need for computer-assisted AAC design and describes a form which computational tools might take.

1.3.4 A Novel Technique for Characterizing Mode Strategies

The dissertation provides a technique for characterizing (qualitatively and quantitatively) the differences among an interlocutor's various *strategies of mode use* — that is, the strategies an individual may potentially use when producing multimodal communicative actions. The technique focuses on the mode strategies of individuals who use Voice Output Communication Aids, which include both *aided* and *unaided* mode strategies. These two types are distinguished on the basis of whether the VOCA is used or not.[5] Previous researchers have identified and analysed various outcome attributes with respect to aided communicators' communicative actions; some of these attributes concern the aided communicators' use of mode(s). The dissertation provides a decision-theoretic formulation in which multiple, relevant attributes of mode strategies are identified and synthesized, thus allowing the merits of the various mode strategies to be evaluated relative to one another. Previous approaches to characterizing the communicative actions of individuals who have communication disorders have presupposed a message-passing model of communication and have focused on the "bandwidth" of the unaided and aided modes. In doing so, these approaches have focused on the individual's *modes of communication* (or at least a subset of them). By contrast, the focus of the characterization technique is on the *repertoire of mode strategies* that is afforded by these modes of communication.

This characterization technique is needed if we are to take into account a crucial fact about the context in which AAC devices are actually used: AAC devices are used *during* the course of a face-to-face communicative exchange. (In fact, the device serves *to mediate*, albeit partially, the communicative exchange.) The dissertation identifies a duality with respect to AAC devices: they are simultaneously devices with which users must interact (and thus, the user's interaction with them needs to be evaluated), and they are also a component of a system that serves to mediate

[5]This term denotes that the individual's repertoire of modes is augmented through the availability of a communication aid. This term does not imply that the use of the mode of synthesized speech — the *aided mode* that is afforded by the device — is necessarily a component of all of the mode strategies in his or her repertoire (e.g., an individual may elect to use his or her other, *unaided* modes, either in isolation or in combination).

communicative exchanges (and thus, the user's contributions to that communicative exchange, both mediated by the device and unmediated, need to be considered).

If the aided communicator employed solely the mode of synthesized speech, then the human–computer interaction aspects of AAC devices would subsume the aspects that concern mediated communication. The rationale for multimodal interfaces for AAC devices would be compelling. However, interlocutors make use of a variety of mode strategies in their repertoire, and, as this dissertation shows, adding multimodality to the interface of the AAC device might affect that repertoire in subtle ways. This dissertation identifies the fact that there are the tradeoffs between the aided mode strategy of synthesized speech and the other multimodal-aided strategies and demonstrates, through example and formal analysis, that improvement with respect to the aspect of bandwidth in the human–computer interface may come at the expense of the mediated communicative exchange (e.g., by virtue of an overall deleterious effect on an interlocutor's repertoire of mode strategies). By considering the mediational aspects of AAC devices, the dissertation exposes the hidden costs of multimodal interfaces.

In sum, one challenge of this work has been to determine the effects of AAC devices in terms of their effects on interlocutors' repertoires of mode strategies (rather than in terms of their effect on a particular mode strategy: the isolated use of the mode of synthesized speech). The characterization technique provides an answer to this challenge.

1.3.5 Computational Instantiation of the Characterization Technique

This dissertation also presents a computational instantiation of the characterization technique and demonstrates the utility of the technique in a novel computational simulation tool that demonstrates the hypothesized relative merits of the various strategies of mode use that are available to a given individual. The characteristics of the VOCA and the individual's capability with respect to the other, unaided modes are parameters to the tool. The characteristics of the device are manipulated, and the consequences of various possible designs are demonstrated through computational simulation. Although simulation tools have been developed to model certain types of human–computer interaction (e.g., for use as an aid for the design of application interfaces), none have been developed for human–VOCA interactions (e.g., for use as an aid for the design of AAC systems). The need for such simulation tools is a logical consequence of the increasingly apparent need for a more principled approach to AAC system design [Light, 1999].

The computational tool is evaluated to the extent that is presently possible, given the currently-available empirical data. The evaluation establishes that the characterization technique does indeed account for the known advantages and disadvantages of various strategies of mode use.

1.4 Outline of the Dissertation

Chapter 2 provides a definition of the modes of communication and an argument that these modes can only be used by participants, collectively, in joint action. The theoretical consequences of this definition are described (essentially, the notion of a mode of communication cannot be used to characterize an individual's autonomous action, and the notion of a mode of articulation must be

defined and used instead). This chapter also provides a definition of the channels of communication and distinguishes them from the modes. It also provides a discussion of multimodal synergy and a criterion for its identification.

Chapter 3 provides an overview of Augmentative and Alternative Communication systems and situates them as but one type of clinical intervention for communication disorders. This chapter argues that AAC systems in fact *mediate* communicative exchanges (rather than merely serving as a prosthesis for the voice) and describes some of the characteristics of AAC-system-mediated communicative exchanges. The process of the design of AAC systems is discussed and analysed — the analysis distinguishes between the evaluation of an AAC system's *local* and *global* effectiveness (e.g., with respect to effect on the individual's repertoire of mode strategies). The chapter concludes with a proposal to use computational simulations to gather information about an AAC system's global effectiveness.

Chapter 4 describes in detail the bottleneck reduction hypothesis, which is essentially that multimodal AAC devices will be more effective than unimodal ones. In this chapter, an analysis of the rationale for multimodal interfaces for AAC devices is presented; this analysis demonstrates that multimodal AAC devices, while possibly improving the utility of the aided mode in isolation, might have a detrimental effect on an individual's other mode strategies. The chapter provides a formal model of the relationship between an individual's resources for communication and his or her repertoire of mode strategies. The chapter also identifies several outcome attributes of mode strategy use that are relevant to aided communicators and describes a multi-attribute model for characterizing an interlocutor's repertoire of mode strategies with respect to the performance of a particular type of communicative action. The technique for characterizing the mode strategies in an interlocutor's repertoire is based upon this multi-attribute model.

Chapters 5 and 6 describe a computational instantiation of the characterization technique. The instantiation is situated as a component of a multimodal surface realization module in the architecture of a communicative agent. The agent architecture is defined as part of a simulation package, MSIM, which is also described in the chapter. The package is used in simulations of multimodal communicative agents engaged in joint activity with one another.

Chapter 7 describes the use of MSIM to investigate the global impact of the effect of changes in an individual's communicative resources on his or her repertoire of mode strategies. These communicative resources may include the use of an AAC device. The chapter describes the parameter value assignments for several simulation conditions. The results for each of the conditions are presented and discussed. The qualitative evaluation demonstrates that the characterization technique successfully captures the tradeoffs between the advantages and disadvantages of each of the agent's available mode strategies. The qualitative evaluation also demonstrates that the characterization technique is sensitive to the features of the current communicative scenario.

Chapter 8 describes aspects of this dissertation that require additional research, and closes with a summary of the dissertation.

Chapter 2

What is Meant by a Mode of Communication?

2.1 Overview

The focus of this chapter is defining and discussing the notion of a mode of communication, and distinguishing the modes of communication from the modes of articulation, the modes of sensory-perception, and the channels of communication.

Section 2.2 will define the term *mode*: a mode, according to lexicographers, is a manner in which a process is performed or is carried out. Thus, the use of the term implies that there is a process for which different manners are being identified and distinguished. (The misuse of the term *modality* is discussed in section 2.2.2.) Thus, the phrase *modes of communication* should refer to the manners in which the process of communication can be carried out, the phrase *modes of articulation* to the manners in which the process of articulation can be carried out, and the phrase *modes of sensory-perception* to the manners in which the process of sensory-perception can be carried out. The modes of communication should not be equated with the modes of articulation (nor the modes of sensory-perception) because the process of communication cannot be equated with the processes of articulation (nor sensory-perception). The section will argue that a definition of the modes of communication should start from a characterization of the process of communication.

Section 2.3 will provide an overview of Clark's [1996] characterization of communication as an emergent process of joint activity and relate this model to information-based models and Speech Act Theory. The definition of the modes of communication will start from Clark's characterization. This section will argue that the modes of communication must be manners of acting jointly. Section 2.4 will discuss the notion of a channels of communication.

2.2 The Term *Mode*

2.2.1 Basic Definition

Subsequent sections will discuss and analyse the advantages and disadvantages of the use of synthesized speech by individuals with physical disabilities. Synthesized speech is produced with a Voice-Output Communication Aid (VOCA), which will be further described in section 3.3.5. In these future discussions, we will show that synthesized speech can be considered to be a mode just like any other — gesture, facial expression, vocalization — albeit a mode with special properties. It will be described as an *aided* mode, whereas gesture, facial expression, and vocalization (among others) will be referred to as *unaided* modes. In the following sections, these modes will be described further; here we explicate the notion of a mode further.

According to one sense identified by lexicographers, a *mode* of some activity or process is a *manner* in which that activity or process is carried out.[1] The term *mode* is used in a wide variety of contexts: modes of travel (e.g., the manners in which one can travel, such as taking a train, bus, or airplane), modes of operation (e.g., the manners in which a machine or device might be operated, such as a digital camera in snapshot or in video mode), modes of payment (e.g., the manners of paying for something, such as cheque, cash, or money order). Note that according to this definition, a mode is inextricably connected to a process — a *mode* is not simply a mode but a *mode of some process.* Also note that the identification of a *repertoire of modes* with which a process can be carried out serves *to characterize* the process.

2.2.2 An Inaccurate Use of the Term *Modality*

Lexicographers define the term *modality* as the *property* of being modal — i.e., modality is a property *of* a process or activity. For a particular process, if manners can be distinguished in which that process can take place, then that process has the property of having modes and thus has *modality.* In the subsequent discussions, the term *modality* will be used this way rather than to refer to any of the modes themselves. This practice differs from the research literature, in which the term *modality* is often used in the latter way. For example (with emphasis added):

- "The term *modality* ... refer[s] to the sense by which information is perceived" [Dannenberg and Blattner, 1992, p. xxiv]
- "a *modality* is a process for analyzing and producing *chunks of information*" [Martin et al., 2001, p. 2]
- "her performance in the oral *modality* was difficult to assess so her [performance] was investigated in the written *modality*" [Brédart et al., 1997, p. 212]

[1] According to the *Oxford English Dictionary* [1989, sense 4 (a)], a **mode** is "a way or manner in which something is done or takes place; a method of procedure in any activity, business, etc." According to *Webster's Seventh New Collegiate Dictionary* [1967, senses 1 (a–c)], a **mode** is "a manner, way, or method of doing or acting: modern modes of travel"; "a particular form, variety, or manner: a mode of expression" or "a given condition of functioning; status: The spacecraft was in its recovery mode".

- "*Modality* refers to the type of communication channel used to convey or to acquire information. It also covers the way an idea is expressed or perceived, or the manner an action is performed." [Nigay and Coutaz, 1993, p. 172]

We might surmise that, in assertions such as those above, the term *modality* was erroneously considered to be synonymous with the term *mode*, and the authors treated them as interchangeable.

2.2.3 Lack of Consensus

Several researchers have commented on the lack of consensus on the meaning of the term *mode* (e.g., Dannenberg and Blattner [1992]; Edwards [2002]). This lack of consensus has been noted both *within* particular research communities (e.g., Human–Computer Interaction and Augmentative and Alternative Communcation) and *between* research communities. An analysis of this lack of consensus can provide additional insight into the notion of a mode and a reconcilation of differences.

The lack of consensus has two possible explanations. One is that researchers simply don't agree with the basic definition given here: that a mode of an activity or process is a manner in which that activity or process is carried out. Alternatively, consensus may exist with respect to the basic definition and the lack of consensus, instead, concerns which process or activity is to be characterized by the defined modes. The determination of which of these two explanations applies is difficult, since in most publications in which the term *mode* is used, an explicit definition is not provided, nor is the process or activity that is being further characterized explicitly identified. A careful reading of the literature suggests that it is the second explanation is the correct one. For instance, *speech* and *writing* are identified as modes in aphasia studies — they distinguish among the manners in which an individual's language production facilities may be selectively affected — and *speech comprehension* and *reading* are also identified as modes — they distinguish among manners in which language comprehension takes place (see [Swindell et al., 1998]).[2] In computer-assisted instruction, *text* and *graphics* have been identified as modes — they are the manners in which the system can provide instructive material to the user. In analyses of the communicative strategies of bilingual speakers, the different *types of language used* (e.g., spoken or signed) have been referred to as modes — they distinguish among the manners in which the speaker could engage a communication partner [Griffith, 1985]. In human–computer interaction, the manners in which human users can perform input actions, such as *keyboard* and *mouse*, are referred to as modes, and the manners in which the system can respond to the user, such as *text*, *graphics*, and *auditory cues*, are also referred to as modes. (These two processes — the provision of information to the computational system from human user, and the perception of information by a human user from the computational system — can be conflated. For example, a mode "is seen as a process for analyzing **and** producing *chunks of information*" [Martin et al., 2001] (emphasis added).) In these examples, the characterizations of a mode are consistent (as manners in which a process or activity is carried out), but the processes that have been characterized differ. The modes that are identified in one context of use need not correspond to those in another, since the processes that they serve to describe and to characterize might differ.

[2] Swindell et al. [1998] actually used the term *modality*, but the assumption has been made here that the term *mode* was intended. See section 2.2.2.

2.3 Defining the Modes of Communication

2.3.1 Overview

It follows from the basic definition of a mode that was given in section 2.2.1 that a *mode of communication* **must** refer to a manner of communicating. Thus, the definition of the modes of communication depends on a characterization of the process of communication itself.

In the following subsections, the process of communication will be distinguished from the related, yet distinct, processes of articulation, cognition, and sensory-perception (which, although involved in the process of communication, are not equivalent to it). Thus, the modes of communication are related to, yet distinct from, the modes of articulation, of cognition, and of sensory-perceptual processing.

A characterization of communication will also provide the basis for a model of dysfunction in communication, to explain the ways in which the process of communication can be affected by dysfunction (e.g., the chronic occurrence of misunderstandings, or the inability of interlocutors to advance the satisfaction of their joint goal). Such models are important, since they provide the rationales behind clinical interventions for communication disorders (to be described further in chapter 3). Despite their importance, they are yet insufficiently developed [Light, 1999].

2.3.2 Interposed Elements

Whether the process of communication occurs face-to-face or through "technological media", it always involves *interposed elements* that mediate between individuals [Craig, 1999, p. 126, 143]. These interposed elements are produced by and perceived by the participants and form the "currency" of communication. The interposed elements also serve as a *de facto* definition of the *context* of a communicative process — i.e., context is everything that these interposed elements are not.

The production and the interpretation of these interposed elements have been characterized in different ways, and each characterization typically has its own term for the interposed elements. They have been variously identified as *messages* [Beukelman and Mirenda, 1998], *speech acts* [Austin, 1962; Searle, 1969], *conversation acts* [Traum and Hinkelman, 1992], *dialogue acts* [Bunt et al., 1995], *conversational moves* [Power, 1979], and *signals* [Clark, 1996]. Various models of the communication can be distinguished by their characterization of the ways in which these interposed elements are produced and interpreted.

2.3.3 The Information-Theoretic Model of Communication

In the simplest model of communication, which has been termed *the conduit model* [Reddy, 1979], these interposed elements are seen as functioning to "transfer" messages from a source (the sender) to a destination (the hearer). The model characterizes communication as an information process going on between at least two human interlocutors [Berge, 1994].[3] By *information process*, we mean any process in which information is transmitted and received. This characterization has its roots in information theory — individuals "encode" and "transmit" messages (i.e., data packets) by performing observable actions and "decode" messages by the processes whereby he or she attends to

and observes his or her environment, forms percepts, and draws inferences from them. A system of turn-taking serves to organize the participants in alternating roles of sender and receiver [Sacks et al., 1974] (a participant is the sender when he or she has the conversational turn; speech overlap occurs when one "transmits" messages during the other's conversational turn).

Information processes occur when individuals simply attend to and observe their environments. If these environments happen to include other individuals, then an information processes among individuals takes place, even when the co-present individuals perform actions without any intended meaning. Thus, one problem for the information-theoretic model of communication is that it over-generalizes. In communication processes proper, speakers *mean things* for their addressees.

Model of Dysfunction

The information-theoretic model of communication is highly prevalent in the Speech-Language Pathology literature. Many clinical interventions for individuals with communication disorders are characterized as providing alternative means for an individual "to compose and to transmit messages" (e.g., see Beukelman and Mirenda [1998]; Lloyd et al. [1997]). This characterization implies that, without the clinical intervention, the individual has little ability, if any at all, to compose and to transmit messages. This view is reinforced by the strong similarity of *reauditorization*[4] to *echo-checking*.[5] This view has also given rise to the notion of *communication rate*, which is used to characterize the "speed" at which an individual is able to transmit his or her messages. Communication rate is typically measured in words per minute (wpm). But a simple count of the words an interlocutor produces per minute, if not correlated with number or duration of his or her conversational turns, can be meaningless. Also, the number of words an interlocutor produces is not necessarily correlated with the effectiveness of his or her communicative action. Furthermore, the quantification of the number of words an interlocutor produces ignores the contribution of other modes that might modulate the meaning of the words, such as facial expression or gestures of the hands, head, or torso. The topic of communication rate is revisited in section 3.3.7.

This model of dysfunction suggests that clinical interventions should provide an alternative facility for message production [Arnott et al., 1988; Copestake, 1996; Todman and Alm, 1997; Vaillant and Checler, 1995]. This approach has often been characterized as providing a *voice prosthesis, speech prosthesis*, or *communication prosthesis*. The approach is predicated on the assumption that, just as a prosthesis can be crafted for a missing limb, so too can a prosthesis be crafted for a missing voice (or for a "missing" apparatus for producing communicative actions, in general). This assumption is problematic, however. As will be described in the next chapter, individuals who have communication disorders typically still use their voices, albeit not for fully intelligible speech; they use them for vocalizations and other sounds. So these so-called prostheses do not actually serve as

[3] In the subsequent discussion, the assumption will be made that communication is a process that takes place between *at least two* individuals. The scenario in which an individual "communicates" with him- or herself is not considered to be a communicative process, but rather a type of information or cognitive process that takes place within an individual's own mind. In addition, subsequent discussion focuses on interpersonal communication and is not meant to apply to other types of communication, such as mass communication or film.

[4] *Reauditorization* refers to the strategy in which the communication partner repeats his or her understanding of the aided communicator's utterance once it has been produced [Bedrosian et al., 1992]). See p. 30.

[5] *Echo-checking* refers to a style of acknowledgment used in computer communication.

replacements; instead, they serve *to afford synthesized speech*, which complements an already-existing repertoire of manners which can be used to perform communicative actions.

2.3.4 Ways of Meaning

Grice [1957] observed that there are crucial differences between the way a person *means* something to another person and the way signs (or events or other entities) *mean* something. Clark [1996, p. 126] describes the difference as one between the meaning of "deliberate human acts" (such as uttering speech or making gestures, which he calls *signals*[6]) and the meaning of "certain natural events" (such as spots appearing on skin as a result of measles, which he calls *natural signs* or *symptoms*). Grice referred to these types of meaning as *non-natural* and *natural* meaning, respectively. According to Grice, non-natural meaning can be further divided into two types, depending on the entity that is the agent of the action of meaning something. In the first type, the speaker S means something for another person A that p (where p denotes a proposition), where S does so by presenting some deliberate human action (i.e., a signal, s) to A. In the second type of non-natural meaning, it is the deliberate human action (i.e., the signal s) that means something, where that meaning can be paraphrased and denoted by a proposition p. These two types of meaning are connected to one another, since a speaker can mean something only through the use of signals, and signals mean something only because they are used by speakers to mean something [Clark, 1996, p. 128].

The formulation of what it means for a speaker to *mean something for another* has been the focus of considerable attention. This is not suprising, since a formulation of speaker meaning is tantamount to the characterization of the process of communication. One formulation that is faithful to Grice's original idea, but has also been amended to reflect the contributions of subsequent arguments, is as follows [Clark, 1996, p. 129–130]: In presenting s to audience A, a speaker S *means for* A that p if and only if the following criterion is satisfied: that S intends (i.e., has intention i) in presenting s to A that A recognize that p in part by recognizing that intention i (i.e., the intention that S intends in presenting s to A...). The criterion has a circularity — the intention that S possesses contains a reference to the intention itself. Thus, the crux of a speaker's meaning is a special type of intention — one that cannot be discharged without the audience's participation.

Clark [1996, p. 30] used Levinson's notion of *activity type* to distinguish between two types of contexts within which actions[7] might be performed by an individual: as part of an autonomous activity or as part of a joint activity. A wide range of activities are joint — two individuals playing a piano duet, paddling a canoe together, playing catch, conducting a business transaction, negotiating an agreement, gossiping, and so on. An *autonomous* activity (or "solo" or individual activity) has a single participant, whereas a *joint* activity is carried out by two or more participants working collectively. This is the essential quality of joint activities — that they require *coordination*. These two types of activity provide two different contexts in which actions might be performed. An action

[6]Clark characterizes the *signals* [p. 155] that interlocutors produce as not necessarily consisting of acts of speech. They may make use of any type of language, spoken or signed (e.g., American Sign Language). Clark assumes a very broad definition of language, which includes many different types of actions, including facial expressions, co-verbal gestures, posture shifts, and others. (These actions might otherwise be considered to be nonlinguistic, paralinguistic, or composites consisting of both linguistic and nonlinguistic components.)

[7]Clark [1996, pp. 18–19] used the term "action" to refer *both* to single acts and to sequences of actions (i.e., activities). To distinguish between the two senses, the term *action* will be used to refer to single acts or deeds, while the term *activity* will be used to describe sequences of actions.

performed as part of an autonomous activity is an *autonomous* action, whereas an action performed as part of a joint activity is a *participatory* one. Unlike autonomous actions, the performance of a *participatory* action requires coordination among the participants.

Clark [1996, p. 130] argues that the performance of the actions in communicative processes are actually participatory actions (also described as joint actions). This assertion is at odds with our intuitive understanding of the acts performed during communicative exchanges as autonomous actions (e.g., acts of speaking, gesturing, and so on). This characterization of communication is also at odds with the information-theoretic model of communication, in which the actions of encoding, transmitting, and decoding messages are autonomous.

Clark's characterization of communicative actions as joint actions has an important consequence for the study of multimodal communication. If the use of a mode of communication can only be achieved through joint action, then it *must refer to a manner of acting jointly*. Just as an individual alone cannot communicate as an autonomous activity, an individual cannot "use" a mode of communication. To characterize an individual as "using" a mode of communication is equivalent to characterizing him or her as "using" a particular manner of communication. But since the process of communication is incontrovertibly a joint activity, then a manner of communication can only be used by the participants jointly.

2.3.5 Speech Act Theory

Austin [1962] observed that an individual's action of speaking, under favourable conditions, may actually *do* something, such as marrying a couple, making a bridge bid, or getting someone to pass the salt, to quit smoking, or to change his or her mind. Some acts of speaking have their effects, under favourable conditions, by changing the mental states of others, and these mental states produce an effect in the things that that people subsequently do. Austin proposed a set of mechanisms whereby the utterances of a speaker might do this. Prior to Austin's proposal, utterances were seen primarily as linguistic entities that had descriptive value — as spoken versions of written statements that make assertions about the state of the world; thus, utterances were often analysed with respect to whether they reflected the *true* state of the world or not.

Austin showed that the actions produced by an interlocutor in a communicative exchange can be analysed from three different perspectives. First, there is the physical action, which can be observed by others and even measured empirically (e.g., the words used and how they are spoken). Using Clark's terminology, the physical action is an autonomous action. Second, there is how the physical action has been perceived by the interlocutors in terms of its perceived function (e.g., whether the spoken utterance is perceived as a declarative statement, a request, or some other type of utterance). Last, there are the consequences that follow from the percepts that the addressees have formed, in terms of modification to their mental states (e.g., whether the addressee actually became convinced of something) and/or to their behaviour (e.g., whether the addressee actually does what was requested). Thus, any particular spoken utterance has effects or facets at these three levels.

Austin originally asserted that for each single utterance, three different acts are performed (which is another way of saying that a single utterance has three different facets):

- The **locutionary act**, which is the act of saying something. The act involves the articulation of speech sounds, which includes the use of words and syntactic constructions (and where those linguistic entities are used with a particular sense and reference). For a locutionary act to be performed, an interlocutor must have the means (such as speech-sound articulators) for doing so.

- The **illocutionary act**, which is the act performed *in* saying something. An act *in* saying something has a functional dimension. Examples of an act *in* saying something include asking, telling, suggesting, greeting, and so on. Illocutionary acts do not come into existence simply by virtue of being performed by a particular individual, but rather are *generated* by locutionary acts under certain conditions. In Clark's terminology, the performance of an illocutionary act is a joint action. The necessary and sufficient conditions for an illocutionary act "to be performed" (jointly) rely upon *both* interlocutors and will be discussed below.

- The **perlocutionary act**, which is the act that is performed *by* saying something. A perlocutionary act refers to the effect a speaker achieves in the listener, such as impressing them, persuading them, or embarrassing them. While it is often written that perlocutionary acts are *performed*, it is probably more accurate to say that they are *produced* on the basis of illocutionary acts and the listener's response to them. (This particular component of Speech Act Theory has been the focus of criticism; see [Marcu, 2000; Gu, 1993].)

In Austin's account, "communication" — the process whereby speech results in people actually doing things — takes places through the effects of illocutionary acts. Searle [1969] provided an account of the necessary and sufficient conditions for the performance of illocutionary acts. This account includes many conditions; we will focus here on those that involve the mental states of the interlocutors. According to Searle [1979, p. 30–31], the basis for determining whether an action is communicative (i.e., whether an illocutionary act has taken place or not) depends upon interlocutors acting with and recognizing intent. For an illocutionary act α to have been performed:[8]

- The speaker must intend to produce in the hearer the knowledge that the illocutionary act α has been made to her or him.

- The listener must recognize the speaker's intention to have performed α.

Model of Dysfunction

According to Speech Act Theory, in order for an illocutionary act to be performed successfully, the speaker must act with intent, and the listener must recognize that intent. Thus, we can model dysfunction in communication on the basis of the ways in which illocutionary acts can be unsuccessful:

- An interlocutor, when in the role of speaker, incorrectly infers which illocutionary acts are likely to bring about desired perlocutionary effects (this inability might otherwise described as a shortcoming in rhetorical skill).

[8]Note that the criteria below are simply another formulation of what it means for a speaker to *mean something for another*; see section 2.3.4 for related discussion.

- An interlocutor, when in the role of speaker, performs locutionary acts from which listeners are unable infer that the illocutionary act α has been made to her or him.
- An interlocutor, when in the role of listener, fails to recognize the speaker's intention to have performed α.

The successful performance of illocutionary act α requires the joint action of the speaker and his or her communication partner (i.e., the listener); each participant must fulfil his or her role. If *either* participant does not fulfil his or her role, the act will be unsuccessful.[9]

Speech Act Theory has had very limited use in the analysis of dysfunction in communication in the Speech-Language Pathology research literature. An exception is the work of Light et al. [1985b], who analysed the types of illocutionary acts that were produced by children who have communication disorders. They defined an *illocutionary act* to be the "speakers' communicative intent or function" [p. 98] (e.g., request for information, provision of information, and getting attention [p. 99]; a complete inventory of the types of possible illocutionary acts was provided by the authors). In their study, a panel of judges analysed a corpus of videotaped interactions; for each interaction, the panel identified the communicative actions performed by the children and coded them according to type of illocutionary act. It was found that children used a different set of illocutionary acts with familar communication partners than with unfamiliar ones.

In order to perform an analysis that is based on the identification and classification of illocutionary actions, a fundamental issue must be addressed. A speaker's illocutionary intent is an internal mental state; it cannot be inspected directly and can only be inferred from his or her surface-level behaviours. For this type of analysis, it is necessary to demonstrate some evidence that the inferences are valid. In Light et al.'s study, high levels of inter-judge and intra-judge reliability were found, but they did not investigate the degree to which the judges' perceptions were consistent with the perceptions of the communication partner (for instance, as might be inferred from the partner's subsequent actions), nor did they state whether the coders made use of information about the partner's reaction in order to code the illocutionary actions.

2.3.6 The Contribution Model

Speech Act Theory provides an account of how spoken utterances, even when highly similar in their physical characteristics (e.g., same words and intonation), can be performed in different communicative contexts, with different intended meanings, and those various meanings can be successfully interpreted. The basic mechanism is the addressee's successful interpretation of the intended illocutionary intent. Allen [1994] provides an example: the locutionary act of uttering "Do you know the time?" might correspond to several illocutionary acts, such as the act of asking a yes–no question, the act of asking for the time, or the act of offering to tell the hearer the time. In order for a speaker to ask for the time successfully, the listener has to recognize that it is this question is being asked, and not the other possible illocutionary acts.

[9] Strictly speaking, if either participant does not fulfil his or her role, the act will be not be performed. Therefore, the act will not even exist (let alone be unsuccessful); extant acts are successful by definition (they are successful by virtue of their very existence, since they are brought into existence *only* when they are successful). Instead, we might say that the *attempted act* was unsuccessful. Note that an attempt implies some sort of originating force; in this case, the origin of the attempt is the intent of the speaker. The previous condition concerns the speaker's ability to "orginate" appropriate attempts.

Although Speech Act Theory connects the utterance of some words and the actual result arising from that action, it does not explain the connection that individual actions have to one another — e.g., as contributions to an overarching, cooperative activity rather than "a succession of disconnected remarks" [Grice, 1975, p. 45]. That speaking and listening are parts of a collective activity (as opposed to separate, autonomous activities) is a tenet in the study of language use [Clark, 1992].

Clark [1996] characterizes communication as a processes whereby shared knowledge is established and argues that this can be achieved only through joint activity. Thus, communication is *an emergent process* of an underlying joint activity. Clark characterizes most, if not all, joint activities as having a *dominant goal* (or *domain* goal), which is established at the outset of the participants' interaction. (Other types of goals are also established; these are described in more detail in section 5.5.)

Empirical evidence shows that interlocutors engage in a process whereby they strive to reach a state of mutual, shared knowledge of what was intended by the speaker of an utterance — this process is described as *grounding* [Clark and Schaefer, 1989]. The hearer of an utterance cannot infallibly recognize the mental state of the speaker and, knowing this fact about the hearer, the speaker cannot be sure that he or she has been understood as intended [Traum, 1994, p. 2]. Once mutual understanding has been established, the *common ground* of the participants accumulates. Establishing mutual understanding does not mean that the speaker and the listener must establish that the listener's interpretation of the speaker's intent is exactly identical to the actual intent. Rather, the degree to which each participant's understanding is shared *needs to be sufficient* for the needs of the communicative exchange at that moment. The degree of sufficiency can vary both within and between joint activities. Clark's Contribution Model [1996] describes both the sufficiency criteria and how they are satisfied. The interlocutors advance toward the joint activity's dominant goal by accumulating common ground.

In order for interlocutors to add to their common ground, they employ *devices for coordination*. It is from this process of coordination that communication *emerges*. Clark identified three such basic coordination devices: *describing-as*, *indicating*, and *demonstrating*. *Describing-as* is a mechanism for *meaning something for another* that is based on the use of a system of symbols whose meaning is shared by the interlocutors, such as systems of spoken or signed languages, or gestures that are used as symbols (such as head nods or the thumbs-up gesture). Clark argues that the mechanism of *describing-as* also requires the use of the other coordination devices, *indicating* and *demonstrating*. *Indicating* is a mechanism for *meaning something for another* that creates indices for the entities to which the interlocutors want to refer. Indices are created by instruments, the most obvious ones being the parts of the body that people can orient (e.g., the finger, hand, voice, eyes, head). These instruments must be used within the context of a locative action such as pointing. *Demonstrating* is a mechanism for *meaning something for another* that is based on the use of *depictive* actions. Just as in the case of *indicating*, instruments are needed and people frequently use their bodies as instruments.

Although the coordination devices, in theory, may be used in isolation, they are most often used in combination. In example (1) below, the two-handed gesture is being used in a depictive capacity, the words *I*, *caught*, *a*, *fish*, *this*, *long* are being used to describe-as (each has a meaning in the English language), and *B*'s voice, having been used to utter the words, serves to locate an entity (*B* himself) with the object of the symbol *I*, which is "oneself" (the self that is indicated by the origin of the voice).

(1) B: I caught a fish this long [ ].
[two-handed gesture which demonstrates length]
[Clark, 1996, p. 174]

Model of Dysfunction

A careful review of the Speech-Language Pathology and Augmentative and Alternative Communication literature has failed to find any application of the Contribution Model in the analysis of dysfunction in communication. However, we can see that the model makes several pertinent predictions. First, one source of dysfunction might be the way in which the domain goal is established (e.g., a participant with a communication disorder might be hindered from participating in its establishment). Another potential source of dysfunction is the establishment of mutual understanding (e.g., the communication partner of an individual with a communication disorder might incorrectly believe that common ground has been established, or an individual with a communication disorder might be unable to ensure that common ground has been established).

2.3.7 The Modes of Communication

Recall from section 2.3.4 the argument that if the process of communication is incontrovertibly a joint activity, then a manner of communication can only be used by the participants jointly. Also observe that the use of a coordination device is a manner of acting jointly. Therefore, the coordination devices provide one basis for a definition of the modes of communication — there are three modes of communication, one corresponding to each coordination device.[10]

Recall that the use of a coordination device is a participatory action — each participant has a role to play in its performance (the role of one is to signal[11] and the role of the other is to interpret the signal). In other words, the use of any of the coordination devices requires the use of some means by one interlocutor whereby behaviours are produced so that they can be observed by others, as well as the use of some means by the others whereby those behaviours can be perceived. If we want to tease apart the modes of communication, we must also tease apart the manners in which observable behaviours are produced (i.e., how signals are "made", or how signalling behaviours are "articulated") and the manners in which observable behaviours are perceived and interpreted.

Similarly, a *mode of articulation* refers to a manner in which (potentially communicative) behaviours might be produced, and a *mode of sensory-perception* refers to a manner in which (potentially communicative) behaviours might be perceived. The following section discusses the notion of a *channel of communication*, which provides a context within which these modes can be further characterized.

[10] It is ironic that, although he avoids the terms "multimodal communication" and "mode of communication" altogether, Clark's characterization of communication provides what appears to be the sole basis in the research literature upon which to distinguish the different modes of communication.

[11] "To signal" is meant to include all of the requisite accompanying mental states concerning intent; see section 2.3.4.

2.4 The Channels of Communication

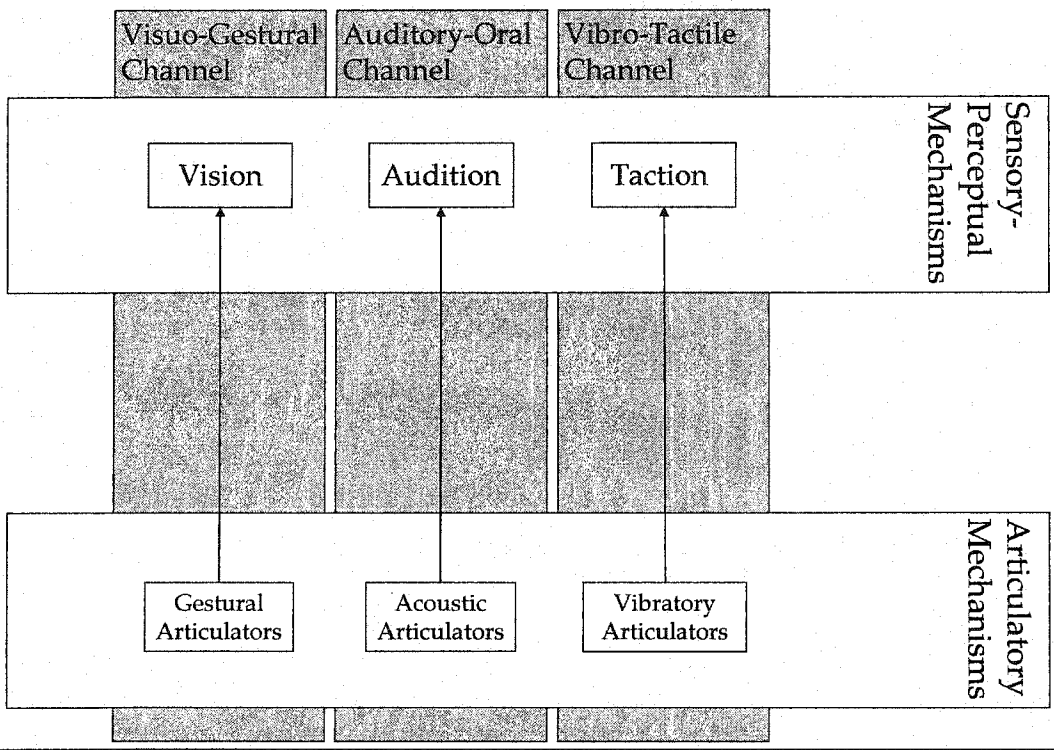
The definition of the *channels of communication* that is used here is based on the notion of the *articulatory-perceptual* (AP) channels — articulatory-perceptual channels are used by human interlocutors when engaged in communicative exchanges [Bouchard, 1996], although not all AP channels are channels of communication. An AP channel is a "conduit" between one interlocutor's means for the articulation of communicative actions and other interlocutors' means for the sensory-perception of the actions that have been articulated. Since interlocutors may have many means for articulation and sensory-perception, multiple AP channels might exist between any two or more participants. Thus, the channels of communication correspond to certain pairwise combinations of the modes of sensory-perception and the modes of articulation.

This definition of the channels of communication differs from others in the research literature. Dannenberg and Blattner [1992] use the notion of a *channel* (when characterizing human–computer interaction) to distinguish among various *sources of information* that are conveyed from the interface to the user (their examples include multiple TVs, multiple acoustic sources, and even the multiple windows on a single display device). In this definition, channels are equated with means of articulation (as opposed to articulatory-perceptual pathways). According to Nigay and Coutaz [1993, p. 172], communication channels are "used to convey or to acquire information." Thus, they characterize the channels of communication both as manners of articulation ("the ways to convey information" *and* as manners of sensory-perception "the ways to acquire information").

Note that information processes (see section 2.3.3) also involves the use of channels. A pair consisting of a mode of sensory-perception and a mode of articulation is a *channel of information* provided that the mode of sensory-perception can be used to perceive actions performed using the mode of articulation. Not all channels of information are channels of communication, however. As described in previous sections, communicative processes involve the perception of intent and the construction of shared knowledge. Thus, channels must be based on manners of articulating actions that not only can be perceived by others, but also *can be understood to have been performed with communicative intent*. A *channel of communication*, unlike a channel of information, must afford to an interlocutor the possibility to act with communicative intent and for that intent to be recognized as such. Vision and audition are the two primary modes of sensory-perceptual processing for communicative actions. Touch is a third possibility; it is the basis for the Tadoma method, which is a communication technique used by individuals with little or no functional hearing or vision.[12] Other modes of sensory-perception, such as gustation and olfaction, cannot serve as the basis for channels of communication, since they lack a plausible corresponding means for the articulation of communicative actions (e.g., acting with intent). Thus, the *auditory-oral* and *visual-gestural* channels are the principal *articulatory-perceptual* channels used by human interlocutors [Bouchard, 1996], and the *vibro-tactile* channel is a third possible channel of communication, albeit one with a highly-specialized use (e.g., for those using the Tadoma method). These channels are illustrated in figure 2.1. The means for articulation and the means for sensory-perception are crucial components of a channel, which is illustrated in the figure.

[12]The technique is based on perceiving speech by feeling the lips, jaw, and throat of a communication partner in order to sense the movements of and the airflow from the lips, as well as the vibrations and movements of the speech articulators) [Lloyd et al., 1997].

Figure 2.1 The articulatory-perceptual channels in human–human communication.



A channel is *open* or *available* from individual A to individual B when A has available the means for articulating observable actions and B has the means for perceiving those actions. A channel is *closed* or *not available* when either of these components is missing. The *direction* of a channel need not be symmetric — a channel might be open from one interlocutor to another, but not necessarily in the other direction. For example, the visuo-gestural channel is open from a individual who is blind but can sign or gesture to an individual who can see, but not in the other direction, and the auditory-oral channel is open from an individual who speaks but is deaf to an individual who hears, but not in the other direction. For communication to take place, at least one channel of communication must be open, in each direction, between the interlocutors.

The modes of sensory-perception can be distinguished on the basis of the different forms of energy to which humans are attuned — e.g., different types of mechanical stimuli in the cases of the acoustic, vestibular, somatosensory senses; chemical stimuli in the case of the olfactory and gustatory senses; light stimuli in the case of vision [Calvert et al., 1998].[13]

Figure 2.1 provides a basis for distingusihing the modes of articulation: as the manners in which observable actions can be articulated. This approach to defining the modes will be elaborated in

[13]The distinction between sensory and perceptual processes, which is an issue in current research, is not addressed here. Perceptual processes, which involve the formation of percepts on the basis of sensory inputs, can be tightly coupled to sensory inputs, especially in the cases of multiple, simultaneous sensory inputs. Empirical evidence demonstrates that multiple, simultaneous sensory inputs are often processed as a whole, resulting in the formation of fused percepts [Calvert et al., 1998].

Chapter 4. Before this, however, communication disorders and clinical interventions for them will be described in Chapter 3. This chapter will also present Voice Output Communication Aids, and the *aided mode* of synthesized speech that they afford.

2.5 Summary

This chapter presented some of the theoretical issues that concern the definition of the modes of communication. Various definitions of the term *mode* were described, as were three contexts in which the term is used (as in *mode* of articulation, *mode* of sensory-perception, and *mode* of communication). The meaning of the term mode from the meaning from the term *modality* (the two terms are often incorrectly used interchangeably).

The modes of communication were distinguished from the modes of articulation and the modes of sensory-perception — a relevant distinction since the modes of communication must refer to manners of acting jointly and, as such, cannot be "used" by an interlocutor. Instead, interlocutors make use of modes of articulation. A definition of the modes of communication was provided, which makes use of Clark's characterization of communication as an emergent process of joint activity. The channels of communication were defined, which provides the basis for the definition of the *modes of articulation* (to be given in Chapter 4).

Chapter 3

Augmentative and Alternative Communication

3.1 Overview

This chapter will describe Augmentative and Alternative Communication (AAC) systems, in general, and will identify some of the major research issues in their development.

In section 3.2, AAC systems are situated as but one possible component of clinical interventions to communication disorders (i.e., not all clinical interventions involve the use of an AAC system). Section 3.3 identifies *communicative strategies* as the main component of AAC systems. An AAC system is, in the most abstract terms, simply a system of strategies and techniques that the participants of communicative exchanges may employ in order to satisfy their communicative goals. Some of these strategies incorporate the use of a computational device (which, in turn, might entail the use of specialized techniques). With such devices, individuals who have little or no functional speech can, through key presses or other input actions, produce synthetic speech. Although AAC devices have a superficial similarity to "prostheses for the voice", they, in fact, serve to *mediate* communicative exchanges, albeit partially.

Section 3.4 describes the design problem for AAC systems. The parameters of the design problem are the *profile of capabilities* of the user and the *target set* of communication scenarios in which the AAC system is meant to be used. For an AAC system to be truly effective, it must be effective in all, and not just some, of the communicative scenarios identified in the target set. A particular feature of AAC design is iteration: effectiveness cannot easily be achieved otherwise. Thus, a crucial ingredient in ensuring effectiveness through convergent design is useful feedback. This feedback, unfortunately, is difficult to obtain using existing practices. One solution to this problem — the use of computational simulations — will be proposed in the subsequent chapter.

3.2 Augmentative and Alternative Communication Interventions

Augmentative and Alternative Communication interventions are clinical interventions that compensate for a particular *pattern of disability*. The term *pattern of disability* refers to any sort of *functional* limitation [RTC/IL, 2001; NCDJ, 2002; DSS-C, 1991]. A limitation is functional when it adversely affects the individual's ability "to perform socially defined roles and tasks within a sociocultural and physical environment" [Beukelman and Mirenda, 1998, p. 241]. Thus, a functional limitation, in general, is one that affects an individual's ability to accomplish his or her goals — each role to be filled or task to be performed is associated with some goal or goals. Functional limitations contrast with physical impairments; physical impairments need not affect an individual's abilities in a way that limits the achievement of their goals.

AAC interventions are intended to compensate for disability patterns that relate to communication, and not other types of disability patterns such as those that relate to mobility (these have their own types of clinical interventions). In other words, the functional limitations that AAC interventions address are those limitations that concern the satisfaction of goals that require communicative exchanges. The previous chapter showed that communicative exchanges emerge from underlying joint activity. By making use of Clark's characterization of communication, we can conclude that the functional limitations that AAC interventions actually address are those limitations that concern goals that are achieved through joint activity.

Notice that this characterization of communication-related disability patterns avoids the term *communication disorder* and, thus, differs from that given by the American Speech-Language-Hearing Association (ASHA). ASHA states that AAC interventions attempt to compensate for "...the disability patterns of individuals with *communication disorders*" [ASHA, 1989, p. 107]. ASHA also characterizes individuals with *communication disorders* as those "...for whom gestural, speech, and/or written communication is temporarily or permanently inadequate to meet their communication needs" [ASHA, 1991, emphasis added, pp. 9–10].[1] ASHA's elaboration on the nature of a communication disorder is circular: "individuals with severe communication disorders are those who may benefit from AAC..." [ASHA, 1991, p. 10].

The notion of a communication disorder is paradoxical. The ASHA definition states that a communication disorder is something that is experienced by an individual — i.e., if an individual experiences inadequacies with respect to communication, then the cause relates to the disorder(s) that affect(s) his or her ability "to communicate." But, as described in the previous chapter, the ability to communicate can be possessed only by groups of interlocutors, collectively. So the ability to communicate is not something that an individual can have, yet an individual can have a communication disorder.[2]

[1] The wording of this definition has been slightly altered for the sake of clarity. The rationale for the modification is discussed in appendix A.1, page 139.

[2] The speech-language pathology research literature tells us that the severity of an individual's communication disorder depends not only on the degree of their primary disorder(s) with respect to motor, cognitive, linguistic and/or sensory-perceptual processes, but also on his or her communication partner. In other words, some patterns of primary disorders result in a disorder of communication with some partners, whereas other patterns of primary disorders do not. Furthermore, for some individuals, a particular pattern of primary disorders might result in a communication disorder, whereas for other individuals, the same pattern does not. Thus, the manifestation of a communication disorder is determined by factors that are not solely dependent upon the individual. Instead, it is more accurate to say that a communication disorder is a secondary disorder that manifests itself in the context of joint activity.

The ASHA definition also characterizes an individual as having communicative needs and identifies them as not being adequately met. But the definition should be generalized to include the goals and joint activity more generally. Communicative exchanges arise from joint activity, and joint activities are established to satisfy a variety of goals (not all of them necessarily associated with the needs of the individual with the disorder).

Communication-related patterns and other patterns of disabilities are intertwined. For example, disability patterns that are related to mobility may affect the degree to which an individual can initiate joint activity in the first place. Although AAC interventions primarily concern compensating for communication-related disability patterns, they must be devised in consideration of other patterns of disabilities

Functional limitations may arise as a consequence of *impairment* (although not all impairment results in functional limitation) [RTC/IL, 2001; NCDJ, 2002; DSS-C, 1991]. The term *impairment* refers to disorder(s) in an individual's primary processes (i.e., an "abnormality of a body mechanism"). The matter of determining the relationship between disorder(s) in an individual's motor, cognitive, linguistic, and/or sensory-perceptual processes and the functional limitations that he or she experiences can be complex, because functional limitations arise in the context of an individual's interaction with his or her environment, which includes other individuals. Factors that concern the environment and other individuals can compound impairment. To say an individual has a functional limitation can be a mischaracterization if it is not qualified — an individual has a functional limitation *in a particular context*.

The consequences of communication-related functional limitations can be serious, such as reduced independence and self-determination, and reduced ability to participate in the community, to be part of social groups, to be gainfully employed, and to receive an education [Blackstone and Pressman, 1995].[3] One proposed, yet unimplemented, approach to the evaluation of AAC interventions is to determine the degree to which the intervention brings about improvements in the individual's participation in the above activities [Beukelman and Mirenda, 1998]. Notice that the activities are joint activities — again, by making use of Clark's characterization of communication, we can conclude that individuals who have communication-related functional limitations, in fact, experience obstacles to their participation (or attempted participation) in joint activities.

Clark argues that conversation — characterized as "the free exchange of turns among two or more participants" [Clark, 1996, p. 4] — should be viewed as the "cradle of language use" [p. 9]. He assumes a very broad definition of language, that includes many different types of actions, including facial expressions, co-verbal gestures, posture shifts, and others — essentially, any action performed as part of joint activity. Conversation is the setting for children's acquisition of language and it is universal to human societies. Other settings of language use, such as lectures, speeches, news conferences, performances, letters, and newspapers, should be seen as altered derivations of the more-basic setting of face-to-face conversation. In face-to-face conversation, participants share the same physical environment (Copresence); not only can participants see and hear one another, but they can do so without perceptible delay (Visibility, Audibility, Instaneity); participants formulate and execute their actions in real time (Extemporaneity), and these actions are fleeting (Evanescence); a participant can perceive and produce at the same time, and the participants can do

[3]This description has been slightly modified for clarity.

so simultaneously (Simultaneity). Thus, another approach to the evaluation of AAC intervention is to determine the degree to which the intervention brings about improvements with respect to the conversations in which an individual participates.

When an AAC intervention is warranted, it is typically devised collaboratively, by an *intervention team.*[4] Such a team consists of professionals whose specialties relate to various aspects of AAC (e.g., speech-language pathologists, occupational therapists, physicians, rehabilitation engineers, among many others), as well as other stakeholders (e.g., the individual with the communication disorder, and his or her parents, friends, and caregivers).

AAC interventions are often multifaceted [Beukelman and Mirenda, 1998]. They may include the removal or reduction of barriers to communication, adaptations to an individual's environment, and further development of an individual's natural abilities through training. An intervention might also include the use of an *augmentative and alternative communication system.* An AAC system is simply a system of strategies that can be used by interlocutors when engaged in, or attempting to engage in, communicative exchanges. Some, but not all, of these strategies involve the use of communication devices. System development involves initial design, implementation and customization (e.g., of the hardware and software of the device, if any), training, evaluation, and modification on the basis of feedback.[5] As will be described in section 3.4, the process of development can be lengthy and complex, and consequently, might not necessarily produce the most effective AAC system for a given individual.

The intervention team considers several factors when developing an AAC intervention: the barriers to communication that presently exist for an individual (and whether they can be removed or reduced); the individual's profile of capabilities and the potential for improvement of his or her natural abilities; the features of the individual's environment that concern his or her ability to engage in communicative exchanges (and whether adaptations might be made); and whether the use of an AAC system is warranted [Beukelman and Mirenda, 1998]. Many factors must be considered together when determining the best overall, multifaceted set of interventional actions.

3.3 Augmentative and Alternative Communication Systems

3.3.1 The Component View

AAC systems are typically described in terms of their components (e.g., see Beukelman and Mirenda [1998]; Lloyd et al. [1997]; Berstein [1988]; Glennen and DeCoste [1997]; Shames et al. [1998]). The four components are:

[4]The criteria for determining whether an AAC intervention is warranted or not can be a source of inequity. In the past, individual organizations (such as clinics or school boards) determined whether individuals were candidates to receive clinical intervention. As a result, the criteria they developed frequently were idiosyncratic, inconsistent, and biased toward individuals with certain types of disorders; many who might have benefited from clinical interventions did not receive them [Beukelman and Mirenda, 1998, p. 146]. However, in Canada, the United States, and many other countries, an individual's right to clinical interventions for communication-related patterns of disability is now mandated by public policy. Consequently, the criteria for candidacy became a matter of public policy as well. This approach has mitigated previous inequity.

[5]It is worth noting that even though most AAC systems are explicitly designed by an intervention team, they might also emerge holistically as well (e.g., they might evolve among an individual and his or her frequent communication partners, such as family members and caregivers).

- **Strategies**: the specific strategies that the interlocutors employ in their communicative exchanges.
- **Communication aid** (also referred to as **AAC device**): a physical object or device that an individual uses.
- **Symbol set**: a set of symbols that is used during the communicative exchange.
- **Access techniques**: the manner(s) in which an individual refers to the elements of the symbol set. The primary kinds are *direct* and *indirect*.

These components form a system. The interlocutors use the strategies during the communicative exchange, and these strategies can entail the use of a communication aid, which entails the use of symbol set and access techniques for that set. These four components can be identified — in at least some form — in every AAC system.

3.3.2 Terminology

The following discussions will focus on communicative exchanges among dyads (as opposed to larger groups of interlocutors). An **aided** dyad is defined as a dyad in which an AAC system is used (and it is assumed that only one interlocutor has a communication disorder). Aided dyads can be further categorized according to how **familiar** or **unfamiliar** the interlocutors are with one another. In an unfamiliar aided dyad, the interlocutors are unfamiliar to one another, whereas in a familiar aided dyads the interlocutors are familiar to one another. (An **unaided** dyad is one which neither of the interlocutors has a communication disorder and in which an AAC system is not used.)

In characterizations of aided dyads, the term **AAC system user** is often used to describe the aided communicator (e.g., see Bedrosian et al. [1992]). It is true that the aided communicator uses the AAC system, and thus, is an AAC system user. However, the term is misleading if applied in the same way the term "telephone user" is misleading when applied to only one participant of a telephone conversation, so we will limit ourselves to the terms **aided communicator** and **unaided communicator**.

3.3.3 Strategies for Communication

AAC systems have as a component a repertoire of strategies for communication. Of the multitude of AAC strategies that have been devised, many concern approaches to producing communicative actions in a way that they are more likely to be interpreted by others as intended.

Many, but not all, AAC strategies require the use of some sort of *AAC device* or *communication aid* (these will be described in section 3.3.5). Some strategies do not involve the use of any external aid. For instance, every interlocutor employs some sort of strategy for turn-taking. But contributions made by aided communicators, whether through existing dysarthric speech capabilities or through using the communication devices are typically produced very slowly and are often unclear, and so conventional turn-taking strategies do not suffice. Additional strategies are often used, such as "expanding out" (or "reauditorization" [Bedrosian et al., 1992]), which involves the communication

partner repeating their understanding of the utterance once it has been produced.[6] This strategy is useful for detecting misunderstandings early. Misunderstandings can be very difficult, if not impossible, to correct once several communicative turns have elapsed. Thus, the overhead entailed by the strategy is justified by its benefits (the overhead is typically — but not always — unwarranted in conversations among individuals without communication disorders). This strategy also allows the communication partner to introduce elaborations into conversation.

Other strategies include role-playing and prompting [Beukelman and Mirenda, 1998], as well as those that target other aspects of participating in a communicative exchange, such as initiating and concluding the exchange; taking, holding, and yielding the conversational turn; correcting misinterpretations made by others. Familiar communication partners generally have more knowledge of the various AAC strategies and have more expertise of when to make use of them. Aided communicators also employ strategies for interacting with unfamiliar communication partners who tend to revert to certain strategies, such as attempting to guess the aided communicator's utterance without allowing him or her to finish.

The aided communicator makes use of all four components of the AAC system: the communication aid, the access techniques, the symbol set, and the strategies. The communication partner primarily makes use of one component, the strategies. Thus, the strategy component of an AAC system is used by both interlocutors in a dyad. Thus, by extension, the whole system is used by both interlocutors, and, in fact, serves to *mediate* the communicative exchange. For this reason, the terms "AAC system user" and "AAC user" doesn't serve to distinguish between the aided communicator and his or her communication partner.

3.3.4 Maintenance and Generalization of Communication Strategies

At this present stage[7] of the AAC movement, the methodology for establishing effectiveness is done by comparing one intervention to another (as opposed to comparing an intervention to none at all).

The primary goal of an AAC system is that it be *effective*. The effectiveness of an AAC system should be distinguished from its *efficiency*. The former relates to the degree to which the system has the desired or intended effect, whereas the latter relates to the amount of effort spent to bring about that effect. (The issue of efficiency will be discussed further in section 6.5.) Notice that in the previous discussions, the goal of an AAC system has been characterized as the goal of *mediating communicate exchanges*. The outcome that is relevant to this goal is the *joint behaviour* of the participants. This contrasts with the characterizations by others, for whom the relevant outcome is defined solely in terms of the aided communicator's behaviour. For instance, Light [1988] has

[6]Note that it is the communication partner who must initiate the use of this strategy. AAC interventions can target not only the individual with a communication disorder, but also his or her communication partners (such as family members, friends, caregivers, teachers). While such strategies have been shown to be beneficial, they alone cannot suffice (it is not possible even to identify all of an individual's potential communication partners, let alone instruct them in advance). There are limitations to the strategy of focusing interventional actions on communication partners.

[7]The clinical use of AAC interventions in a particular municipality, province or nation is typically mandated by health care policy, and these policies typically progress through a well-known sequence of stages [Beukelman and Mirenda, 1998] (which gives rise to an AAC "movement"). In the initial stages of an AAC movement, the merit of AAC interventions themselves must be established, usually through case studies in which the benefits achieved from AAC interventions are demonstrated by comparison to scenarios in which interventions were missing. Once the movement advances, the methodologies for evaluation (which demonstrate the effectiveness of AAC interventions) become more rigorous.

argued that the criteria for evaluating effectiveness should be based upon the demonstration of *communicative competence*, a construct which requires an aided communicator to have skills from each of the four "domains" of linguistic, operational, social, and strategic competence. Others share this view, that the evaluation of effectiveness should be based on measuring the system's effect on specific intrapersonal behaviours (i.e., the degree to which the behaviour of the aided communicator has modified by the introduction of a communicative strategy):

- "*Effectiveness* refers to the demonstration of behavior change as a direct result of treatment" [Schlosser and Braun, 1994, p.208].
- A change in behavior is the presumed effect or consequence of the intervention [Schiavetti and Metz, 1997].
- The dependent variable in efficacy research is the target behavior that is to be changed by the AAC intervention [Light, 1999, p.14].
- In Schlosser and Lee's [2000] meta-analysis of 54 AAC evaluation studies, the outcome attributes that were used almost exclusively focused on the aided communicator's behaviour.

Light [1999, p. 14] argued that the targeted behaviour must have *social validity* — it must be considered by the participants and other relevant stake-holders to be an important behaviour that will legitimately serve to enhance the communication of the participants.[8]

With this approach to evaluation, the effectiveness of an AAC system gets established once the aided communicator demonstrates the acquisition of certain, desired intrapersonal behaviours. However, the communication partner's behaviour plays a significant role in creating the opportunities for the aided communicator to demonstrate these behaviours. Also, a change in one participant's behaviour might not necessarily bring about a change in the outcome of the joint activity. This confounding might explain the findings of Lasker and Bedrosian [2001], who showed that evaluations of a particular AAC intervention as both positive or negative depends on the criteria applied. The aided communicator's behaviour should not be the only attribute upon which evaluation should be based.

The designers of AAC interventions (at least those who characterize the "effect" of intervention in terms of strategy acquisition by aided communicators) strive to demonstrate that an individual who has been taught to use a particular communication strategy maintains and even generalizes the use of that strategy. Treatment effects might not be maintained; they might be seen in the short term, but be lost over time. For instance, an interlocutor may initially use a particular strategy, but gradually cease to use it. Schlosser and Lee [2000] identified an inventory of strategies for the promotion of generalization and of maintenance which included: "train and hope", "train to criterion and hope", "environment modification", "reduction of discriminability of consequences", "train to generalize", "positive reinforcement (of unprompted generalizations)", and several others. During the maintenance condition of an evaluation study, if it has one, an assessment is made of the degree to which previously-identified intervention effects can still be identified. Maintenance phases typically employ a single follow-up probe or multiple follow-up probes design.

[8] See appendix A.2 (p. 139) for a note concerning the wording of this statement.

The designers of AAC interventions also strive to demonstrate that the effects of the intervention are generalized. Effects might be specific to examples used in training, and not generalized to other applicable situations. During the generalization phase of an evaluation study, if it has one, an assessment is made of the degree to which intervention effects can be found, if any, in "conditions" (i.e., communication scenarios) other than those that were explicitly targeted during training. Generalization phases typically employ either a single-, multiple-, or continuous-probes design.

Because AAC systems are highly tailored to the specific needs of each individual, a *within-subjects* design must be used — the same outcome variables are measured with respect to the same dyads, but under different conditions (such as pre- and post-intervention). The use of a *between-subjects* experimental design — an experiment in which the outcomes from a group of different subjects in a given condition are synthesized and compared to other such syntheses — is ill-advised because the pool of individuals who use AAC systems is small and highly heterogeneous.

Because so few subjects can be solicited for any particular study, a small-n experimental design must be employed, such as an AB, ABA, ABAB design or some other configuration of A and B phases. The A phase is when baseline observations are made. This is typically followed by a phase during which the AAC system is introduced, and training and instruction is provided to the users.[9] The B phase is when the effects of the AAC intervention are observed. The derivation of the baseline provides a quantitative standard against which the effect of an intervention can be measured, but some evaluation studies omit this phase. These studies are now of limited use [Schlosser and Lee, 2000]. Other evaluations in the AAC literature have been anecdotal, and did not make use of controlled experimental conditions.

3.3.5 AAC Symbol Sets

A prevalent view in the AAC literature is that the behaviour of individuals in a communicative exchange can be characterized as sign production — spoken, written, or gestural. Signs are carriers of meaning and are often identified as either linguistic (e.g., words) or nonlinguistic (e.g., facial expressions). For an individual who has a communication disorder, the repertoire of signs that can be produced (at least without intervention) is constrained by the capabilities of his or her own body. These signs often are few in number, cannot be interpreted by others (e.g., dysarthric speech), and do not correspond to a shared system of meaning (e.g., certain signs can be produced and reliably reproduced, but they are not the signs that are already associated with conventional meanings). Each of these problems has its own remedy.

One remedy is to establish, in advance, the meaning of those signs that an individual is able to produce using his or her own body, such as gestural signs that are formed using hand shape, sequences of eye blinks, or movements of the legs or head. (Such signs will be described as *interlocutor-producible* in subsequent discussions) Through the conventionalization of the meaning associations, which must be acquired by both the aided communicator and his or her communication partners, these signs become symbols. The set of signs for which meanings are assigned is termed the *AAC symbol set*. If the communicators are literate, then signs that correspond to the

[9] The AAC system is considered the "treatment variable" — a term borrowed from the case study approach to behavioural change. AAC interventions are often characterized as treatments, and the outcome that follows from interventional action is seen as the treatment's effect.

Figure 3.1 A communication board in use. Image used with permission, Mayer-Johnson (2004). The Picture Communication Symbols ©1981-2004 by Mayer-Johnson LLC. All Rights Reserved Worldwide. Used with permission.

letters of their written language might be used. Alternatively, gestural signs might represent what would otherwise be words or phrases. Such schemes have the advantage that the symbols can be produced whenever they are needed. On the other hand, they also require that the conversational participants have advance knowledge of the sign-meaning associations.

To circumvent difficulties associated with interlocutor-producible signs, an approach may be used which includes devising a set of symbols, such as icons, pictures, alpha-numeric characters, written words or phrases. (This set is also termed the *AAC symbol set*.) A tangible display for the symbol set is created (e.g., a physical display, such as a flat piece of cardboard, the pages of a book, or a computer screen). Any such display is referred to generally as a *communication aid*. A communication board (CB) is a common type of communication aid, shown in figure 3.1. With a non-computational communication aid, an individual *refers* to the symbols with his or her body rather than producing them directly.

An alternative to orthographic symbols is the use of icons. Such symbol sets require a scheme for associating meaning both to individual icons and to sequences of icons (i.e., the analog to the meaning associations with orthographic representations). The symbol-meaning associations must be learnable by the conversational participants, as well as being easily recalled. One such scheme, called *semantic compaction*, makes use of one-to-many icon-to-meaning associations [Baker, 1982].[10] Many individuals find the polysemy of icons intuitive. In the example below, different icon sequences, each of which includes the polysemous ⟨APPLE ICON⟩, are given, along with their semantic associations [Conti et al., 1998]:

⟨APPLE ICON⟩ + ⟨NOUN ICON⟩ = "food"
⟨APPLE ICON⟩ + ⟨VERB ICON⟩ = "eat"
⟨APPLE ICON⟩ + ⟨ADJECTIVE ICON⟩ = "hungry"
⟨QUESTION MARK ICON⟩ + ⟨APPLE ICON⟩ = "When are we eating?"

[10] *Minspeak* is the system that is sold commercially that makes use of semantic compaction; the system also includes the approaches for developing and maintaining vocabulary sets.

Externally-represented AAC symbol sets have many advantages over interlocutor-producible symbols. First, the AAC symbol set can be larger and more finely-grained. Meanings that might otherwise be difficult to associate with an interlocutor-produced sign can instead be represented by symbols. The number of symbols that can be referred to is constrained by the physical medium upon which the set is to be displayed, but there are a number of different techniques to accommodate this (e.g., the use of different screens if a computer display is used, or the use of different overlays if a non-computational display is used).

A second advantage is that an externally-represented AAC symbol set can used by a computational device, which can then display them, receive selection input actions, and then link these selection actions to a text-to-speech module. Thus, the device affords the mode of synthesized speech. Such computational AAC devices are described as Voice-Output Communication Aids (VOCAs). From this point onward, we focus exclusively on VOCAs.

A third advantage is that the technique of *scanning* can be used, which is possible only with an externally-represented AAC symbol set. For an individual to make use of any AAC symbol set, an *access technique* is needed. An individual might select a symbol directly by pointing to it or touching it on the display; this access technique is termed *direct* (also referred to as *random*, by analogy to access techniques for computational storage media — e.g., random-access vs. sequential-access). But another technique — called *scanning* — allows individuals who do not possess sufficient motor control to point to or touch a particular symbol on a display to still make a selection from among a set of candidate targets.

In this technique, a symbol is selected through a process in which several choices are made in sequence. The elements of the *selection set* (e.g., typically the set of alphanumeric characters) or a subset of it are displayed in a two-dimensional matrix. In the row–column version of sequential access — widely accepted because it achieves a good compromise between input speed and user ability [Damper, 1984] — a moving cursor highlights successive rows, pausing briefly on each, until an input action is received from the user, such as a button push. Then, the successive columns of the row are highlighted until the next input action is received. A symbol is thereby selected and the process can begin again.[11]

Scanning is a useful scheme for selecting from among a set of symbols, and it is often the only means available to individuals with severe physical disorders. Depending on the nature of an interlocutor's physical disorder, it might not be possible for him or her to press accurately one of many small, closely-arranged keys on a standard keyboard. For some, it is possible only to press a single button (of a customized size and sensitivity) that is mounted exactly next to a particular body part (such as hand, foot, forehead, and so on). Although such devices are severely limited in the type and number of different possible input actions, they are still useful in situations in which an individual must make a choice that has a small number of possible answers.

3.3.6 Voice Output Communication Aids

The aim in designing a Voice Output Communication Aid (VOCA) is "to help users to translate their thoughts accurately into synthetic speech output" and, to accomplish this, two different ap-

[11] If the device is non-computational, then the communication partner must point to each column and row, stopping when the individuals indicates.

proaches can be employed: *phrase-selection* and *phrase-construction* [Todman and Alm, 2003]. With *phrase-selection*, the aided communicator selects (or "retrieves") a string from among a large set of candidates, which is then articulated using synthetic speech.[12] The set of strings must be derived in advance and in anticipation of the future needs of the aided communicator. With *phrase-construction*, the aided communicator composes an utterance by selecting one or more symbols from the AAC symbol set, thereby constructing a string which is subsequently passed to the TTS module. At any given point in a conversation, it is generally faster for an aided communicator to select a phrase than to construct it, but the phrases that are available to be selected are often inappropriate or at least inferior to one that can be tailored to the current context. Because of these complementary strengths and weaknesses, it is desirable for a communication aid, if possible, to incorporate both of these approaches.

A number of techniques have been developed in order to improve phrase-selection. A tendency in early systems was to treat all of the phrases in the set as equiprobable (see [Light, 1989] for a discussion); the same amount of work was required to select any phrase, regardless of the current topic or the phase of the conversation. A number of different models have been developed which allows the VOCA to hypothesize the set of phrases that are most likely to be needed at a given point in a conversation or for a given topic-oriented discussion [Alm et al., 1992b, 1993]. The VOCA then makes these phrases more easy to select than the other possible phrases.

In order to improve the speed at which phrases can be constructed, a number of word prediction and word completion techniques have been developed. For example, after a certain number of letters have been selected, WordQ [Shein et al., 2001] derives a set of words that are candidate completions for the initial string. (Attempts to develop word prediction techniques that exploit information about syntactic part-of-speech co-occurrence, such as by Van Dyke [1991], have not been met with great success in the past, but are the focus of on-going research.) Whichever word-completion or -prediction technique is used, the size of the candidate list (the set of possible next words or completions) must be chosen carefully. The larger the size, the greater the likelihood that the target word is included among the candidates, but the higher the cognitive burden on the user to peruse the list and to make a selection (and, in the case of scanning, the longer it will take for the system to successively highlight each option).

If the scanning access technique is used, the n-gram probabilities can be exploited to minimize the number of binary choices required to select an element of the AAC symbol set. The VOCA can incorporate a dynamic display on which selection set changes between successive symbol selections. With this approach, the probability distributions of letter occurrences in words can be exploited by placing the most likely candidates early in the sequence of presentation. For example, after one letter has been selected, the display of the device might be modified so that those letters that are more likely to follow are more convenient to select (e.g., after the letter "q", the letter "u" is proposed as the next choice).

Whether direct or indirect selection is used, to operate the device and to select elements from the symbol set, the user must perform *some* kind of input action.[13] The interface of a VOCA can be

[12] In *phrase-selection*, synthetic speech includes both synthesized speech, which is generated on-the-fly using a text-to-speech module, and digitized speech, which is recorded in advance and subsequently retrieved.

[13] Although prototype input devices have been developed for use even in situations in which *no* volitional motor movement is possible — e.g., the control of a virtual mouse through the use of neurotrophic electrodes implanted in the motor

customized so that it can make use of whichever types of input actions the user is able to perform. Possible customizations range from binary switches to modified keyboards, or even standard computer keyboards. In fact, customizations can allow essentially any type of action — not only the hands, but also the feet, elbow, shoulder, forehead. Oculomotor activation is possible through the use of an eye-gaze tracker. If a VOCA interface recognizes single, discrete input actions made using a single input device, we describe the interface as *unimodal*. If a VOCA interface recognizes single, discrete input actions, but where those actions may be produced using one of several different input devices, we describe the interface as *multimodal*.

A multimodal interface was developed in an experimental setting that was able to recognize input actions made using a binary switch *or* using vocalization (the latter was achieved through the use of speech recognition module) [Treviranus et al., 1991]. The reported results, although preliminary, indicated that the ability to switch between the two modes of input was beneficial to the user (the number of vocabulary selections per minute increased, when compared to trials in which switches were the only input device permitted). "It is the inherent flexibility of a multimodal input system that is its main advantage. If one particular mode of input is difficult to use, the simple substitution of another ... should alleviate the problem" [Keates and Robinson, 1998]

One possibility for a multimodal VOCA interface — not yet implemented in any system — is for it to recognize multiple input actions as having meanings that are different from the meanings of the individual input actions. That is, the multimodal VOCA interface recognizes the *co-ordinated* production of multiple input actions, each performed using a different device. Each mode is associated with a domain of possible input actions; when two such domains are combined pairwise, the number of distinct input actions is the product of the two domain sizes. Through this mechanism, the use of "a number of modes can increase the vocabulary of symbols available to the user" [Keates and Robinson, 1998]. This possibility motivated the formulation of bottleneck-reduction hypothesis, which is discussed in more detail in section 3.3.8 below and in chapter 4.

3.3.7 Characteristics of AAC-System-Mediated Communicative Exchanges

As described in section 2.3.3, the message-passing model of communication is prevalent in the AAC literature. This model, in essence, holds that AAC systems are used by individuals in order "to compose and to transmit messages" that otherwise could not be composed and transmitted (e.g., see Beukelman and Mirenda [1998]; Lloyd et al. [1997]). This model also has given rise to the notion of *communication rate*, typically measured in words per minute (wpm) (i.e., the rate at which the aided communicator "transmits" his or her messages). Although the notion of communication rate is problematic (see section 2.3.3), the contrast between the communication rates in unaided and aided communicative exchanges is noted here since it is so striking. When an aided communicator employs scanning, he or she typically achieves rates of 0.5 to 5 words of synthesized speech per minute [Vanderheiden, 1988], a rate that is significantly lower than the conversational rate of 150 words per minute or more [Venkatagiri, 1998]. Foulds [1980] found that an individual with no clinically-defined physical disorders could achieve a rate of 8 words of synthesized speech per

cortex [Moore and Kennedy, 2000] — the assumption is made here that the use of VOCAs entails some sort of motoric input action.

minute using scanning, and Yoder and Kraat [1983] report that rates of 5–30 words of synthesized speech per minute can be achieved with direct selection.

Several empirical studies have noted that in AAC-system-mediated communicative exchanges, the aided communicator holds far fewer conversational turns, and the turns that are held are far shorter than those of the unaided communicator [Light et al., 1985b; Beukelman and Yorkston, 1980; Yoder and Kraat, 1983]. (Conversational turns need not necessarily alternate between the participants.) Significant differences between aided dyads and unaided dyads have been reported in repair strategies, feedback responses, and turn-passing strategies [Light et al., 1985c]. Communication partners in aided dyads also may terminate the communicative exchanges prematurely or avoid them in the first place.

Empirical studies have also noted significant differences between the *types* of turns that the aided communicator and the unaided communicators take. Because aided dyads have a predominance of question-answer routines, the unaided communicator's pattern of turns tends to be characterized by a large portion of *yes–no* and specific *wh-* questions, and the aided communicator provides *yes–no* responses and the requested information [Light et al., 1985a]. Light et al. [1985b], analysed the *variety* of illocutionary acts performed by aided communicators in interactions with different types of partners. They argued that most of their subjects used a *smaller* variety of illocutionary acts in the interactions with their familiar communication partners than with unfamiliar ones. (They stated that these results were consistent with those found by Blackstone and Cassatt [1984].) The decrease in variety is somewhat counterintuitive — one would expect that with a familiar partner, aided communicators might have a greater freedom of expression. In addition to being unintuitive, their argument also rests upon results that were probably affected by uncontrolled, and very likely confounding, factors. One uncontrolled factor was the topic or structure of the communicative exchange. The communication partner and the aided communicator were free to structure the communicative exchange as they desired. Possibly, familiar communication partners were more adept at controlling tightly the interactions (e.g., to occupy the "conversational space" and to oblige specific responses). The subjects were not required to produce certain types of illocutionary acts, nor was it ensured that the subjects were afforded the time or opportunity to formulate every possible type of illocutionary act. The difficulty in identifying illocutionary acts on the basis of locutionary ones, which was described in section 2.3.5, also may have also resulted in coding inaccuracy. Until these methodological issues are addressed, it is problematic to assert that aided communicators use fewer types of illocutionary acts with familiar partners than with unfamiliar ones.

A frequently-reported result is that communication with familiar partners is more likely to be *successful* than with unfamiliar partners. Modes of communication such as speech, gesture, and facial expression may be severely affected, but individuals may be able to communicate with family members very effectively using these natural modes, whereas they may need to rely on AAC techniques with unfamiliar partners [Mirenda and Mathy-Laikko, 1989]. Indeed, the verbal mode is used more frequently with familiar partners than unfamiliar [Beukelman and Yorkston, 1980]. Along with being able to interpret imprecise vocalizations, "familiar communication partners are able to visually discern intention from imprecise and idiosyncratic gesture" [Roy et al., 1993a, p. 99].

3.3.8 The Bottleneck Reduction Hypothesis

This section describes Shein et al.'s [1990] proposal for improved AAC devices and analyzes their proposal. The basis for their proposal is an analysis that contrasts *aided familiar*, *aided unfamiliar*, and *unaided* dyads.[14]

Shein et al. point out that even when an aided communicator has little or no functional speech, he or she is able to make use of other modes in *aided familiar* dyads,[15] such as facial expression, body language, vocalizations, eye gaze, and manual or head gestures, *in addition* to the aided mode of synthesized speech.[16] An aided communicator is able to produce multimodal communicative actions, and the partner is able to interpret "nuances of body language and facial expression" [Shein et al., 1990, p. 37]. Not only is a familiar communication partner more likely to be able to successfully interpret an aided communicator's actions, he or she is more likely to be able to infer communicative intent in unmediated actions (rather than attributing the actions to spasticity or other non-communicative causes). Familiar communication partners are able to discern intention from imprecise and idiosyncratic gestures. Aided communicators have multimodal communication skills that are not harnessed by existing communication devices [Light et al., 1985c; Beukelman and Mirenda, 1998].

In *aided unfamiliar* dyads, on the other hand, the communication partner is less able or unable to interpret the aided communicator's unmediated, multimodal communicative actions. The aided communicator's ability to use effectively the unaided modes is constrained by the partner's unfamiliarity, and he or she must rely more heavily (although not necessarily exclusively) on the AAC device to *mediate* his or her communicative actions.[17] Thus, the extent to which the communication channels in an *aided* dyad need to be mediated by the AAC device depends on the familiarity of the communication partner.

Shein et al. argue that the AAC-device-mediated component of the communication channels is hindered by a bottleneck at the interface of the AAC device. The *bottleneck-reduction hypothesis* is that reduction (or even elimination) of the bottleneck will improve the capacity of the communication channels, thereby improving the overall effectiveness of the AAC system.

They argue that there are basically two approaches to interface bottleneck reduction. The first is

[14]These dyad types were described previously, in section 3.3.2: an *unaided* dyad is one in which neither of the interlocutors have a communication disorders, and an *aided* dyad is one in which only one of the two interlocutors has a communication disorder and whose exchanges are mediated by an AAC system. Aided dyads were further distinguished: *aided familiar* dyads, those in which the aided communicator and the unaided communicator are familiar, and *aided unfamiliar* dyads, those in which the aided communicator and the unaided communicator are not familiar to one another.

[15]Note that Shein et al. identified facial expression, body language, vocalizations, eye gaze, and manual or head gestures as different "channels of communication", but their notion of a channel actually corresponds to a mode of articulation, as the term was defined in section 4.2.2. Shein et al. also make use of the notion of "input channels", which will be referred here to as *modes of input* to the AAC device.

[16]Nigay and Coutaz [1993] describe this as a *synergistic* use of multiple modes — that is, two or more modes are used in parallel (with respect to the temporal dimension) and the modes must be interpreted together, as each mode provides incomplete information — see section 6.4.2 for further discussion.

[17]Shein et al. argue that the aided communicator needs to "transfer" more information to the AAC device, so the device can "convey" more information to the communication partner. An alternative way of formulating the argument is that the aided mode needs to *mediate* more effectively the aided communicator's communicative actions (i.e., mediate in the sense of serving to bridge the gap between the input actions that an aided communicator is able to provide and the communicative actions that his or her partner is able to understand.) Thus, we will replace Shein et al.'s information-theoretic formulation of the process of communication with one based on the Contribution Model (discussed in Chapter 2). Instead of distinguishing among different types of *information* being transmitted, we will instead distinguish between input actions to the AAC device and communicative actions performed for the communication partner.

to improve the effectiveness of the input actions that are performed using existing interfaces (which are unimodal). These approaches were described as increasing the *rate* of information transfer through existing input channels. This might include the use of techniques based on language-models, to minimize the number of input actions required by the user. For example, information about character bigrams has been exploited in scanning techniques (see section 3.3.6) in order to reduce the number of input actions required (e.g., by manipulating the order in which selections are presented so that the most likely candidate appears earlier). In addition, information about word bigrams has also been incorporated into word entry, so that the number of input actions required to enter text might be reduced.

The other approach to reducing interface bottleneck is to increase the number of modes of input to the device.[18] Shein et al. note that the interfaces of existing unimodal AAC devices are not capable of capturing modes of articulation other than gesture in isolation (e.g., gaze, vocalization, other types of gesture, such as gesture using the head or torso). Shein et al. argue that these modes represent potential resources that might be exploited for the production of input actions to the AAC device. Note that the modes that they propose to recruit are the same modes that aided communicators would otherwise be able to put to use with familiar communication partners.

Last, Shein et al. point out that the first strategy — to improve the design of unimodal interfaces — is the most commonly-pursued strategy, and they argued that it has only ever resulted in (and only ever will result in) small, incremental improvements. They argued that the drastic and non-incremental improvements that are actually required can be achieved only by the latter strategy and, consequently, that multimodal interfaces for AAC devices should be developed.

The bottleneck-reduction proposal has resonated with other researchers and triggered an initial burst of activity toward the implementation of this proposal. Roy et al. [1993a] sought to develop techniques to distinguish "intention from imprecise and idiosyncratic gestures"; they described a prototype system that included multiple input devices and sensors (including sensors for position, facial expression, and vocalization) and the plan to integrate the multiple streams of sensor/device signals [Roy et al., 1993a]. An evaluation of the implemented system showed that a gesture recognition technique that made use of input both from surface electromyography electrodes (which detect the co-contraction of antagonist/agonist muscle pairs) and from a 3D magnetic arm motion tracker *together* performed better than gesture recognition performed using either mode of input alone [Roy et al., 1994c]. Although Roy et al. [1993a] identified the development of a "new generation of intelligent AAC devices" as a potential use of their multimodal gesture recognition technique, such devices have not been developed and the use of multiple modes of inputs has primarily been to increase recognition accuracy [Roy et al., 1993b].

A number of experiments have been conducted to determine the feasibility of an interface that recognizes the *co-ordinated* production of multiple input actions. Smith et al. [1996] and Dunaway et al. [1986] hypothesized that the capability to recognize such input actions will translate into many benefits to the aided communicator (such as improved speed of input, improved input accuracy,

[18] This was characterized as increasing the number of input channels and as "increasing bandwidth". Shein et al. argued that the aided communicator needs to "transfer" more information to the AAC device, and described the aided communicator's input actions, as well as the unmediated actions that he or she performs for the communication partner, as *information*. They did not identify explicitly the AAC device and the communication partner as receivers of this information, but this is implied in their discussion. See footnote 17 re: the issue of a information-theoretic model of communication.

increased flexibility, and improved "naturalness" of interaction). Smith et al. [1996] and Dunaway et al. [1986] reported the results of pilot studies that determined the types of input actions that users are able to produce using each of the modes of vocalization and head pointing. The next study in their planned sequence was designed to investigate the integrated use of those two modes. Unfortunately, these results were not published.

Keates and Robinson [1998] developed a multimodal interface to a simple system which was controlled by a set of six commands (UP, DOWN, LEFT, RIGHT, YES, NO) and observed its use by six different subjects (all of whom were affected by a motor disorder). For each of the six commands, two input actions were developed; one input action performed using the head and the other using the hand.[19] In one experimental condition, the subjects were shown a pair of different commands and were prompted to produce the input actions corresponding them, one command using the head and the other using the hand. The results showed that, even when given the option of any order of production, virtually all subjects opted to do the head gesture followed by the hand gesture. In the final analysis, the data transfer rate in this condition (measured in bits per second) was lower than in two other conditions (in one, a single input mode was used and, in the other, two input modes were used to signal the *same* command). Keates et al. concluded that the cognitive load on users is "simply too great" when they are required to generate simultaneously gestures of different types. It does not follow, however, that users will not be able to successfully deal with the load under improved conditions. In effect, users were asked to produce multimodal input actions that corresponded to one of 30 possible command pairs[20] that either didn't have a defined meaning in the context of a task (such as "UP-YES") or, worse, had a meaning that was incoherent (such as "UP-DOWN"). Furthermore, the only reward for correct production was on-screen feedback. If the multimodal input actions instead had meanings that were compositional and coherent, and if the correct production of such actions were linked to success in communicative task that is actually needed to perform a day-to-day activity (such as ordering fast food or completing a purchase in a store), then it is reasonable to expect better results.

The ability to recognize coordinated, multimodal input actions remains a promising avenue for the development of improved VOCA interfaces. Only a few studies have been conducted, and their results are either incomplete or inconclusive, so future work is certainly warranted. However, no work to date has acknowledged the issue that multimodal input actions can have a detrimental impact on the other, unaided modes of communication. This issue will be discussed further in chapter 4.

3.4 The Design of AAC Systems

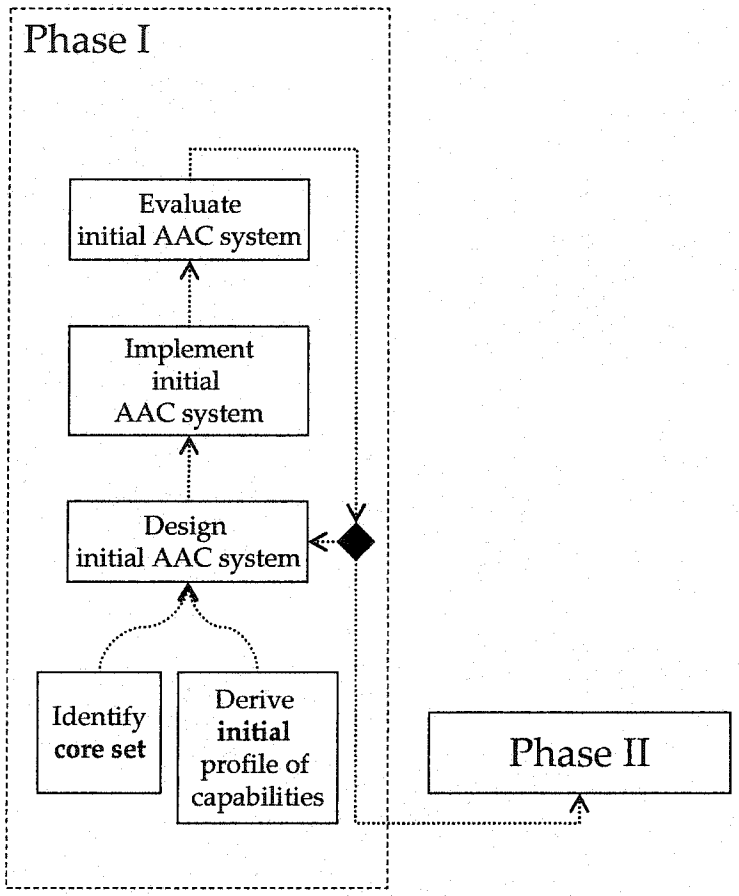
3.4.1 Overview

As the sections below will describe, each AAC system must be tailored to the circumstances in which it will be used, and each set of circumstances presents its own design challenges. An AAC

[19] Head gestures were recognized by a Polhemus tracker, and hand gestures were performed using a joystick of the same type used in computer games or to operate a wheelchair.

[20] There are $6x6 = 36$ possible command pairs, but 6 of these are duplicates (e.g., "UP-UP").

Figure 3.2 An overview of the components of the first phase in the development cycle of an AAC system.



system has essentially two parameters: its user's *profile of capabilities* and the *target set* of communicative scenarios. The design process for AAC systems involves both the identification of these parameter values and the development of a system to suit them.

The process of AAC system designed, which has been characterized as "trial and error" [Light et al., 1990], typically progresses in three main phases [Beukelman and Mirenda, 1998, pp. 149–150]. Each phase can be seen as an iteration on the design of the AAC system. Evaluation is performed during each phase, and modifications are made. Ideally, these modifications will result in successive improvements, and the process will converge on the best possible AAC system for each individual.

3.4.2 Design of Initial AAC System

In the first phase of the AAC design process, an initial AAC system is developed. An overview of this phase is given in figure 3.2.

An aided communicator's *profile of capabilities* characterizes his or her physical, cognitive, lin-

guistic, and sensory capabilities. (The discussion here will focus on physical capabilities, primarily with respect to motor control. Other abilities, such as cognitive and linguistic/language abilities, are also relevant, but will not be discussed here.) In the first phase of AAC system design, an initial assessment is performed. (And in subsequent phases, the individual's capabilities are reassessed and the profile updated.) Motor control relates to the individual's ability to produce speech sounds and gestures and to provide input actions to an AAC device (e.g., typing on a keyboard, operating a binary switch). Thus, an individual's physical capabilities determine both the unaided modes of articulation that might be employed in a communicative exchange and which input actions an individual is capable of providing to an AAC device (the aided mode).

In the first phase of AAC system design, what will be described here as the *core set*, a subset of the communicative exchanges that the AAC system is intended to eventually mediate, is identified. The core set identifies "exchanges of immediate importance" [Beukelman and Mirenda, 1998, p. 149], and includes those required for basic care-giving, for the expression of needs and wants, and for the individual's participation in the development of the AAC intervention. These exchanges typically involve communication partners who are familiar to the individual (e.g., family members and therapists).

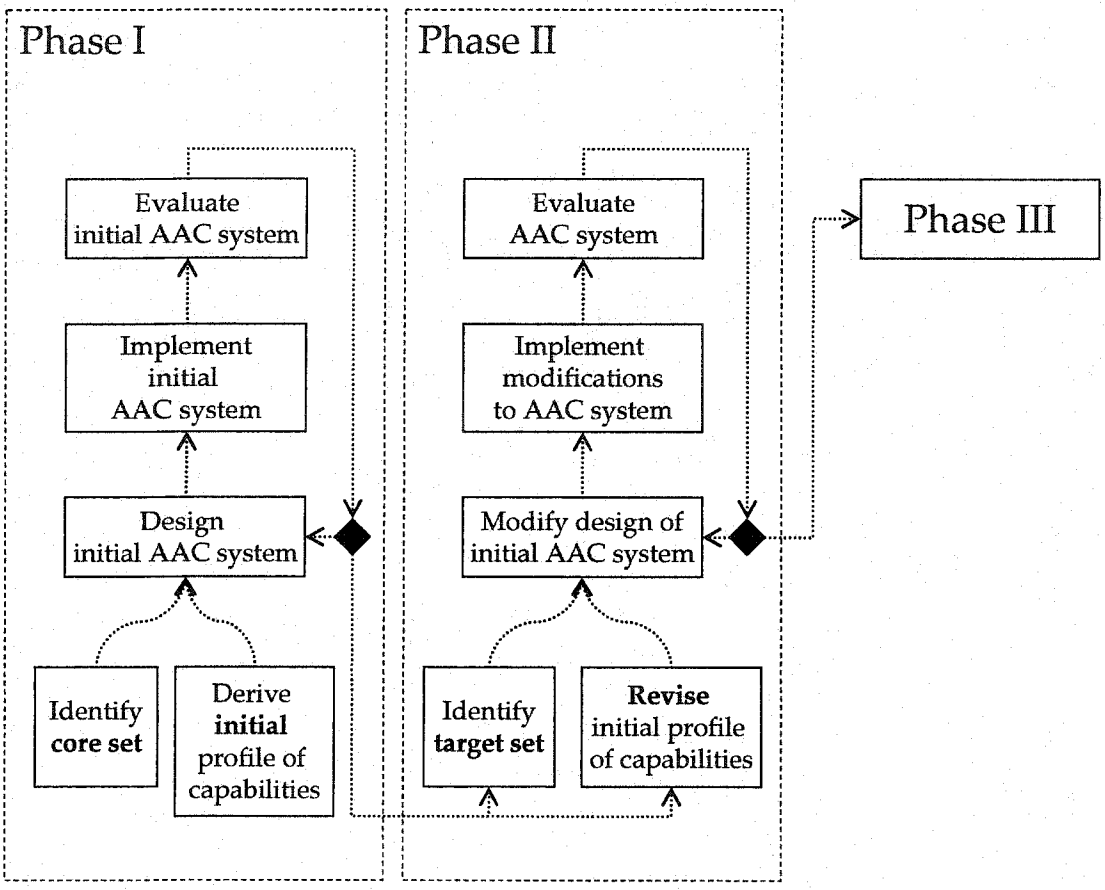
From the initial profile of capabilities and the core set, an initial AAC system is designed. The intervention team can capitalize on the fact that communicative exchanges with familiar partners are easier to mediate than those with unfamiliar partners. Once the required components of the initial system are identified, they are obtained (e.g., by purchase or in-house design) and customized as required (e.g., by specifying the symbol set, by the adding input switches, or by configuring the keyboard). Voice Output Communication Aids, even though commercially available, require additional configuration before they can be used (e.g., vocabulary elements must be selected and programmed and linked to the symbol set). A focus of the first phase is for the aided communicator and his or her communication partners to learn the use of the system and the device. The system will likely be reconfigured in the subsequent design phases, as the individual becomes familiar and proficient with it.

3.4.3 Transition from Initial AAC System

In the second phase of the AAC design process, the design of the initial system is modified and elaborated. An overview of this phase is given in figure 3.3, where the second phase is shown in contrast to the first phase.

In the second phase, the design is extended so that it can mediate in a greater range of communication scenarios than identified in the core set. It is augmented to include other communicative scenarios that are considered important by the aided communicator or the intervention team. This elaborated set will be described here as the *target set*. The target set identifies a wider range of joint activities and larger variety of communication partners (including both familiar and unfamiliar partners). Underlying joint activities can vary widely (e.g., they might include making plans with a friend, making a purchase, or transacting business more generally, requesting information, or making a classroom contribution). Even individuals seemingly engaged in conversation for conversation's sake are actually engaged in a joint activity (e.g., adhering to social norms, fulfilling

Figure 3.3 An overview of the components of the first and second phases in the development process of an AAC system.



social obligations, or building social closeness).

The mediation of all the communication scenarios in the target set presents additional challenges to the AAC system. To meet this challenge, the intervention team must make use of as much of the individual's existing physical, cognitive, linguistic, and sensory capabilities as possible. Thus, a subsequent, more detailed (and correspondingly, more time consuming) assessment is conducted, and the individual's profile of capabilities might be updated.

It is difficult, if not impossible, to identify accurately the target set. To do so entails specifying accurately in advance the way in which an AAC system will actually be used, which entails predicting the types of communicative exchanges an individual will want to or need to engage in, possibly with little or no prior patterns or evidence upon which to base predictions. An accurate specification of the target set is of paramount importance, however, since inaccuracies can mean the AAC system will be used in scenarios for which it was not designed. Thus, the target set devised in the second phase might be revised repeatedly in the third phase of design; each time, the design of the AAC system is modified accordingly.

The type of communication partners the aided communicator will encounter is also important

to the design of an AAC system. Communication partners can vary widely with respect to their knowledge of AAC strategies, their skill in interpretation, and their patience. Depending on the partner, the register of social formality that is appropriate or required may differ. The strategies used by the aided communicator must be designed to suit the full range of partners that he or she will encounter (especially unfamiliar ones). Communication strategies for familiar partners should be devised whenever possible. The context of the underlying joint activity is important to the design of an AAC system. For instance, the vocabulary that an individual needs (and which must be available in the AAC device) is often highly dependent on the joint activity.

In the second phase, the intervention team modifies the initial AAC system on the basis of the (possibly) updated profile of capabilities and the target set. The modified system is evaluated and adjustments, if necessary, are made. The symbol set and the vocabulary of the AAC device may require modification to suit the target set. The design process them moves to the third phase, during which the effectiveness of the design is maintained.

3.4.4 Maintenance of AAC System

In the third phase of the AAC design process, the effectiveness of the system is monitored and the system is modified when warranted. An overview of this phase is given in figure 3.4, where the third phase is shown in contrast to the first and second phases.

The AAC system may become less effective if the aided communicator's capabilities change. Improvements can be made for reductions in physical capability, which often occur in progressive neurological disorders. Alternatively, a gain in proficiency can be exploited by the addition of features to the AAC device. Changes in the individual's abilities must be detected so that the system can be adjusted accordingly; changes in an individual's profile of capabilities change one of the parameters of the design process.

The AAC system may also become less effective after changes in the aided communicator's lifestyle, such as those that relate to his or her stage in education, position in employment, living environment, hobbies, circle of friends, or family situation. Moreover, the intervention team might identify new communication scenarios for the AAC system to mediate, in order to promote social development. In these cases, the target set needs to be updated.

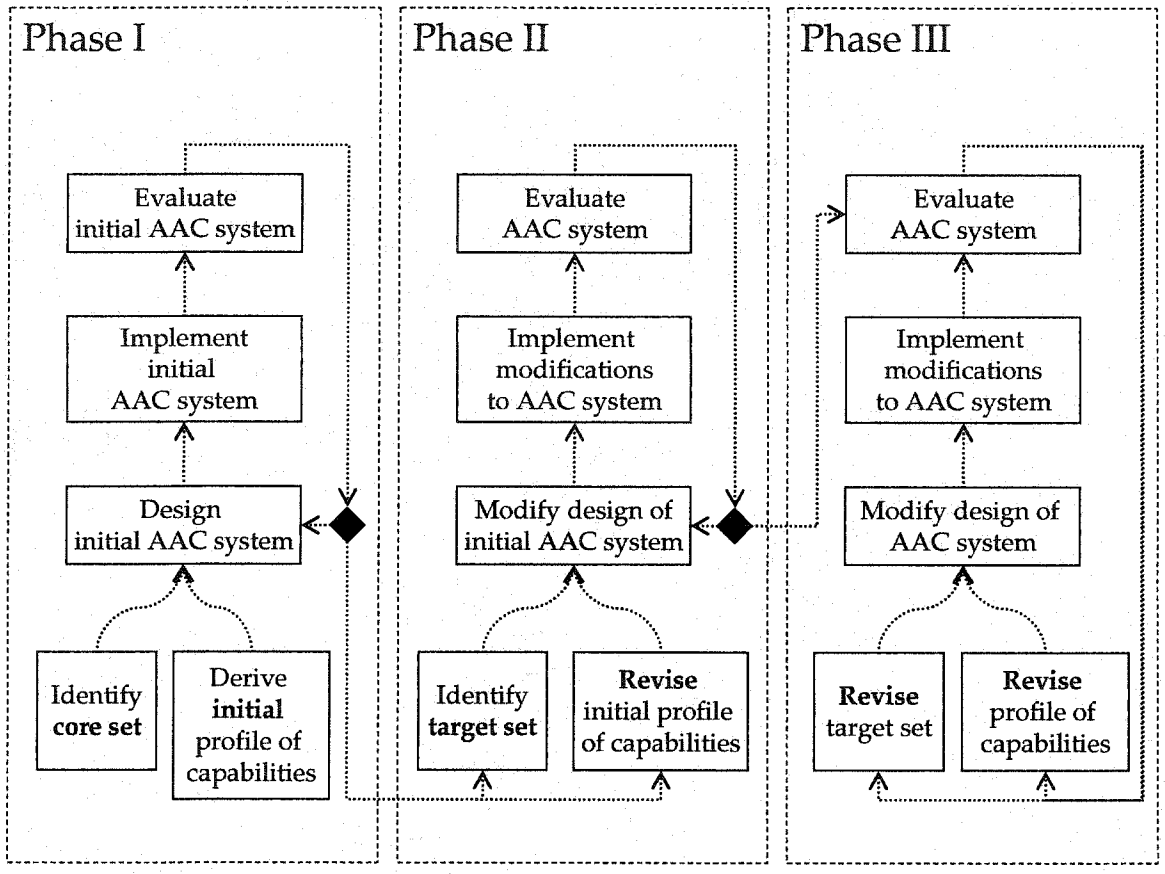
Any time the profile of capabilities or the target set changes, the design of the system may require adjustment in order to match the new specification.

3.4.5 Formal Description and Analysis of AAC System Design

In the preceding sections, a sequence of three phases of the design process was described. This sequence can be seen as one approach to implementing the design process: We now describe the approach in formal terms below.

We define the *parameter space* to be all possible pairwise combinations of all possible profiles of capabilities and all possible target sets. Clearly, the parameter space is huge (in all likelihood, infinite). We define the *design space* to be all of the possible configurations of an AAC system (every possible communication strategy, used in conjunction with every possible AAC device, where each is configured with every possible access technique, every possible symbol set, which is mapped to

Figure 3.4 An overview of the components of the first, second, and third phases in the development process of an AAC system.



every possible vocabulary set). Clearly, the design space is also huge (in all likelihood, infinite). The AAC design task is to identify the AAC system in the design space that best suits the identified profile of capabilities and target set.

The design activities in AAC design are the same as in other design processes. They include: analysis of requirements (i.e., determination of the individual's profile of capabilities and identification of the target set), specification of a system to meet the requirements (i.e., identification of the AAC device and the strategies to be used), implementation of the system (i.e., customization of the AAC device and development of appropriate communication strategies, as well as training), and evaluation. The sequence of three phases that was described in the preceding section provides one approach to performing these design activities. The phases also make use of an iterative approach (e.g., one iteration in each of the first two phases, and subsequent iterations in the third phase).

One important issue for AAC design is that the process requires that the parameter values be accurate. Each AAC system needs to be tailored to the capabilities of the aided communicator, as well as to the communication scenarios in which it will be used. The interface of an AAC device, in particular, must be tailored to the individual's profile of capabilities. Inaccuracies in the profile specification can result in a mismatch between the types of actions the aided communicator is able

to produce and the types of actions required to operate the device. Possible types of inaccuracies are under-estimation and over-estimation of the individual's capabilities. An interface that doesn't fully capitalize on them might require numerous simple input actions where fewer, more-complex ones would suffice. An interface that is too demanding will likely result in a high proportion of erroneous input actions and excessive fatigue. A profile specification might even be inaccurate in both of these ways. This issue can be successfully addressed by evaluation, which reveals any problems that arise from such mismatches. The intervention team, upon discovering this problem through evaluation, would revise their assessment of the individual's profile of capabilities. Moreover, because once-accurate parameter values can become inaccurate, an iterative approach, which includes the periodic re-evaluation of the accuracy of the profile of capabilities, is important.

Another issue for AAC design is the large space of parameter values. The capabilities of individuals with communication disorders can vary widely, and communication scenarios can vary widely as well. There is only a small likelihood that the same exact design parameters will re-occur. If the design of a pre-existing AAC system is to be reused, it would nonetheless require refinement.

Yet another issue for AAC design is the unpredictability of evaluation outcomes. The intervention team might not be able to predict fully the consequences of their design choices. In order to evaluate some aspects of the effectiveness of an AAC system, it must be actually developed and deployed (e.g., to determine the degree to which the aided communicator experiences fatigue with a device, or the effect of the AAC system on the attitudes of the communication partners). An iterative approach to design addresses this issue. In each iteration, the intervention team develops a particular configuration of components and observes the system in actual use in order to gather feedback on the design, then, the design is modified to address its previous shortcomings.[21] Each additional iteration that is undertaken, however, adds time and expense to the development process. For the design process to converge on the best possible AAC system for each individual, each design modification must bring about improved effectiveness and an adequate number of iterations must be performed. But, as will be argued in the following paragraphs, the evaluation of the effect of a design modification can be complex.

First, note that the criteria for evaluation are derived from the design specification, and the design specification evolves over the design process. Thus, the type of evaluations that are done by the intervention team vary according to the phase of the design. For instance, the evaluation done in the first phase differs from that done in the second phase — the first-phase evaluation does not even address the issue of whether the basic AAC system can be used successfully to mediate a novel communicative exchange with an unfamiliar communication partner, since this was not one of its intended uses (i.e., this communication scenario was not in the core set). In the second phase evaluation, however, it would not be adequate to evaluate the modified system solely on the basis of scenarios in the core set.

If we acknowledge that the effectiveness of an AAC system can vary according to the communicative scenario, then a distinction is needed between *global effectiveness* and *local effectiveness*. The latter refers to a system's effectiveness in a particular communication scenario, whereas the former refers to its effectiveness over all of the communication scenarios in the target set. Global effective-

[21] It is also worth noting that this feedback is also needed by those who must justify existing or proposed funding policies on AAC intervention.

ness is more complex to evaluate than local effectiveness because its evaluation must synthesize a set of evaluations of local effectiveness for each of the communication scenarios identified in the target set. It is not clear how this synthesis should be performed (i.e., what relative contributions of each of the evaluations of local effectiveness should be). For instance, they might be weighed equally, by their relative frequencies, or by their relative importance (if measurement of importance is possible). A detailed review of the AAC literature has failed to reveal any such integrative evaluations.

The evaluation of an AAC system's global effectiveness reveals its strengths and weaknesses with respect to the target set. The system may be more effective in mediating some communication scenarios than others. It is also possible that a modification to the system's design might not necessarily just mitigate its weaknesses. It might also reduce its effectiveness others scenarios. Thus, an improvement in effectiveness in one communication scenario may come at the detriment of effectiveness in others. This tradeoff might not be revealed if the local effectiveness of the AAC system is evaluated for only a subset of the target set. Thus, in order to identify design modifications that might inadvertently be detrimental, accurate feedback of the AAC system's global effectiveness is required. Evaluations that provide incomplete or inaccurate feedback can be detrimental to subsequent design and can preclude monotonicity of convergence of the design process.

The evaluation of global effectiveness requires feedback of the AAC system's local effectiveness for each of the targeted communication scenarios. Yet such feedback can be difficult to obtain, since it can be impractical to gather it. The target set might be large and some of the targeted communication scenarios may arise infrequently. (It also may impose on the aided communicator's privacy).

In sum, the evaluation of global effectiveness requires the consideration of the outcomes of a possibly large number of communication scenarios, and each scenario's outcome can be complex. The intervention team endeavours to evaluate the AAC system to the fullest extent that is possible, but, unfortunately, their resources are often constrained (e.g., time and equipment). This can restrict the scope of the evaluations that can be performed. In order to circumvent these problems, the use of computational simulations to gather additional feedback will be proposed in Chapter 4.

3.5 Summary

This chapter provided an overview of Augmentative and Alternative interventions and described the some of the major issues in their development and design. AAC system were described and situated as but one type of AAC intervention.

Two types of AAC systems were described: those which make use of non-computational communication devices, such as communication boards (or no communication device at all), and those which make use of voice-output communication aids (VOCAs). A distinction was made between the user of an AAC device (the aided communicator) and the users of an AAC system (the conversational participants). Because all of the participants make use of the AAC system, it serves to *mediate* communicative exchanges.

Some of the challenges in the design and development of AAC systems were identified and described, which demonstrated that is difficult to design an AAC system from first principles and

that the design process needs to be iterative. A large number of system parameters, each with many possible alternatives, creates a large space of possible AAC systems and, thus, a large space of possible design outcomes. The requirements that each individual has — which are parameters to the design process — can vary considerably, which adds complexity to the process. In addition, although the design process for AAC systems has been performed repeatedly by various clinicians and for various types of user requirements, even already-existing designs must be tailored if they are to be reused. A further difficulty is that several different attributes of the intervention outcome are relevant, and they can be complex, intertwined, and difficult to measure. Not only does the *local* effectiveness of an AAC system need to be evaluated, but these evaluations must be integrated into an overall evaluation of its *global* effectiveness. In the next chapter, the relationship between an aided communicator's repertoire of modes and the global effectiveness of the AAC system will be investigated and a formal model is proposed. In the subsequent chapter (Chapter 5), a computational instantiation of the formal model will be described.

Chapter 4

The Augmented Repertoire of Mode Strategies

4.1 Overview

In this chapter, the problem that is addressed by this dissertation is stated. The starting point for the problem statement is the argument by Shein et al. [1990] that incremental improvements to unimodal VOCA interfaces are not likely ever going to produce the improvements that are actually needed for AAC systems to be truly effective and that we need a way to exploit multiple modes. This chapter identifies a reason why this needs to be done carefully.

In section 4.2, we describe the mechanism whereby a repertoire of modes is afforded both by a set of communicative effectors and by a VOCA. This section introduces the concepts of *articulatory support* and *mode conflict* and relates them to a communicator's ability to use, in sequence or simultaneously, the modes in his or her repertoire.

In section 4.3, we describe the way in which a VOCA with a unimodal interface augments a communicator's repertoire of mode strategies and we contrast this with the anticipated effects of a VOCA with a multimodal interface. The analysis allows us to formulate the problem of interest in this thesis: although a multimodal VOCA will likely afford a more effective aided mode of synthesized speech than a unimodal VOCA, is it necessarily the case that it affords a more effective and productive communicative interaction than a unimodal one? This is the key idea that will be investigated in the subsequent chapters: when comparing and contrasting multimodal and unimodal VOCAs (or any two VOCAs for that matter), the *repertoire of modes strategies* that each type affords should be considered, rather than any particular mode strategy in isolation.

4.2 The Repertoire of Modes and Mode-Specific Sub-Actions

4.2.1 Communicative Effectors

A *communicative effector* is any body structure that generates movement that can be a component of communicative action or that can be an input action to a VOCA.[1] The communicative effectors of human interlocutors include the various musculoskeletal systems for the limbs, hands, head, and torso; the facial effectors (lips, eyebrows, eye, nostrils, mandible); the oculomotor effectors; and the speech-sound articulators, which include the phonatory articulators (larynx), and the resonance articulators (lips, glottis, velopharynx, mandible).

4.2.2 The Modes of Articulation

Recall from chapter 2 that the modes of communication were characterized as manners of acting jointly; they were distinguished from the modes of articulation, which *can* be used by an individual alone. As a starting point, we define the modes of articulation as speech, facial expressions, and gestures of the hand, head, and torso. When the interlocutor uses a VOCA, synthetic speech is yet another mode of articulation.

But what does "using a mode" actually mean? Frege's *principle of compositionality* holds that the meaning of the whole is a systematic function of the meaning of the parts. Typically, the "whole" refers to written sentences, and the "parts" are the individual morphemes, the smallest meaningful unit in the grammar of a language. We hypothesize that this principle applies to multimodal communicative actions as well — that the "whole" of a multimodal communicative action can be decomposed into temporally-coordinated constituents. These components are the multimodal generalization of morphemes — the smallest meaningful units in the multimodal "grammar" of face-to-face conversation. Each sub-action is performed using one and only one mode of articulation. We will use the term *sub-action* to describe these constituents.

Through analysis and decomposition of the multimodal communicative actions of participants engaged in joint activities, inventories of *sub-actions* specific to each mode have been identified. Various systems have been developed for the characterization of spoken utterances, many of which also provide a means for characterizing coverbal gaze (see Schiffrin [1994] for an overview). The Facial Action Coding System (FACS) provides an objective basis for characterizing different facial expressions [Ekman and Friesen, 1978; Ekman et al., 2002]. Early versions of FACS also provided a system for the classification of gestures of the head; others have since been developed (e.g., Kapoor and Picard [2001]). Several researchers have developed systems for the classification of gestures of the hand, and, as pointed out by Thórisson [1996], most of them are modifications of the classification proposed by Efron [1972]. Table 4.1 provides a categorization of the different types of gesture, and distinguishes between deictic, iconic, and metaphoric gestures (these three different types of gestures were used in the simulation, which will be described in section 7.2). As will be described

[1] The term generalizes the term *body effector*, which is used in the speech production research literature. For example, "In order for an individual to articulate a [spoken] language, she must know words of that language, know how to combine words into phrases, and be able to instantiate those phrases in the physical world through the use of *body effectors*" [Byrd and Saltzman, 2002, p. 1076, emphasis added]. (The robotics control research community uses the term *end effector* in an analogous way.)

below, the subset of each inventory of mode-specific sub-actions that a particular communicator is able to produce depends on his or her articulatory suppport.

Table 4.1 One categorization of the different types of gesture, a synthesis of Thórisson's overview and the classifications given by Ekman and Friesen [1969] and Poyatos [1980] (and subsuming that of McNeill [1992]).

1. **Symbolic or Emblematic:** gestures that have a direct interpretation in a given culture, such as the "peace sign".
2. **Deictic:** gestures used to point to a object that is present either visually or symbolically. The hand position of a deictic gesture is often an extended index finger, but not necessarily so and might be performed by any extensible part of the body (e.g., hand, arm, or head) [Quek et al., 2001]. The components of a deictic gesture are the pointing sign itself (the hand position), a *deictic field* (the spatial domain that contains both the intended referent, the addressee, and the speaker), and an *origo*. An origo is the point or perspective from which the pointing originates [Bühler, 1982].
3. **Iconic:** gestures in which the hand (usually) plays the part of another object for the purposes of demonstration (e.g., a cupped-hand gesture to represent a bowl).
4. **Pantomimic:** gestures in which the hands of the gesturer depict hands in another situation (e.g., the gesture that accompanies the word "this" when someone says "he waved bye-bye like this").
5. **Metaphoric:** gestures, similar to iconic, in which the hands depict abstractions rather than objects.
6. **Nondepictive:** gestures that serve as speech markers, such as stressing elements of speech, introducing new elements into the discourse, and regulating turn-taking.
7. **Blends:** gestures that are blends of two or more of the above types (e.g., an interlocutor performing an iconic gesture while outlining an object [Gullberg, 1999], or performing two gestures in sequence without a global rest in between).

4.2.3 Articulatory Support

The use of a mode of articulation requires the support of one or more underlying communicative effector. For instance, the use of the mode of facial expression requires the support of the facial musculature. The degree to which a mode can be used depends on the degree to which the support required from the underlying communicative effectors is available. We believe that the constraints on the underlying communicative effectors determine the types of sub-actions in each mode's inventory. For instance, disorder in the motor processes can affect the neuromusculature of the hands, which may render the mode of gesture wholly or partially unavailable. The disorder has the effect of modifying the inventory of mode-specific sub-actions; it might reduce the number of sub-actions (e.g., a smaller range of gestures) or it might affect their manifestation (e.g., gestures might still be performable, but affected by spasticity or coarseness of movement).

Interlocutors can adapt to reduced or modified inventories of mode-specific actions. For instance, in some settings, idiosyncratic shared systems of meanings might evolve naturally between

an individual with a physical disorder and familiar communication partners. One such idiosyncratic shared system of meaning, which is based on gestures of the leg, was developed by Huer [1987] to circumvent hearing loss and cerebral palsy with severe upper extremity involvement. Other, more-conventionalized shared systems of meaning, such as Signed Exact English, can be established though clinical intervention [Beukelman and Mirenda, 1998]. In cases such as these, the shared system of understanding *mediates* the use of the gestural mode. Thus, even a modified inventory might still be of use to an individual if the remaining mode-specific actions can be successfully interpreted by others.

4.2.4 Articulatory Support from Multiple Effectors

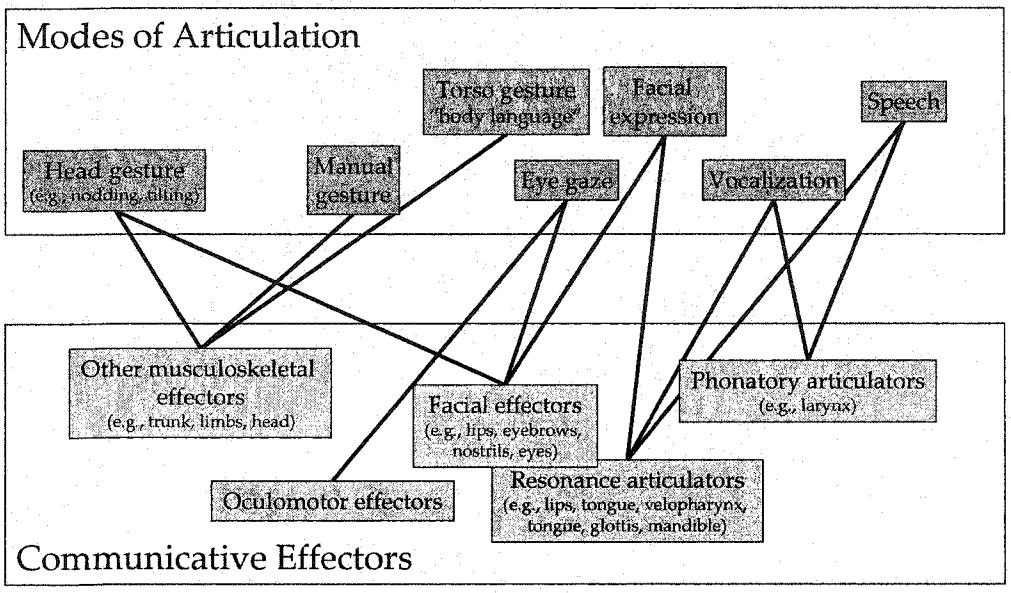
In figure 4.1, the relationship between communicative effectors and modes are shown for two mode repertoires: that of an unaided communicator and that of an aided communicator with VOCA-afforded synthesized speech. The communicative effectors and the modes can stand in many-to-many relationships; that is, an effector can support multiple modes and a mode can require the support of multiple effectors.

For instance, speech requires effectors for phonation and for resonance. Either may be constrained: an individual might have sufficient motor control for the formation of speech sounds, yet have inadequate breath support to cause the vocal folds to vibrate. Conversely, an individual might be able to cause the vocal folds to vibrate, yet not have sufficient motor control of the glottis and lips. Moreover, fluent speech requires that the effectors be coordinated; if there is inadequate coordination, then the mode of speech will be constrained or unavailable.

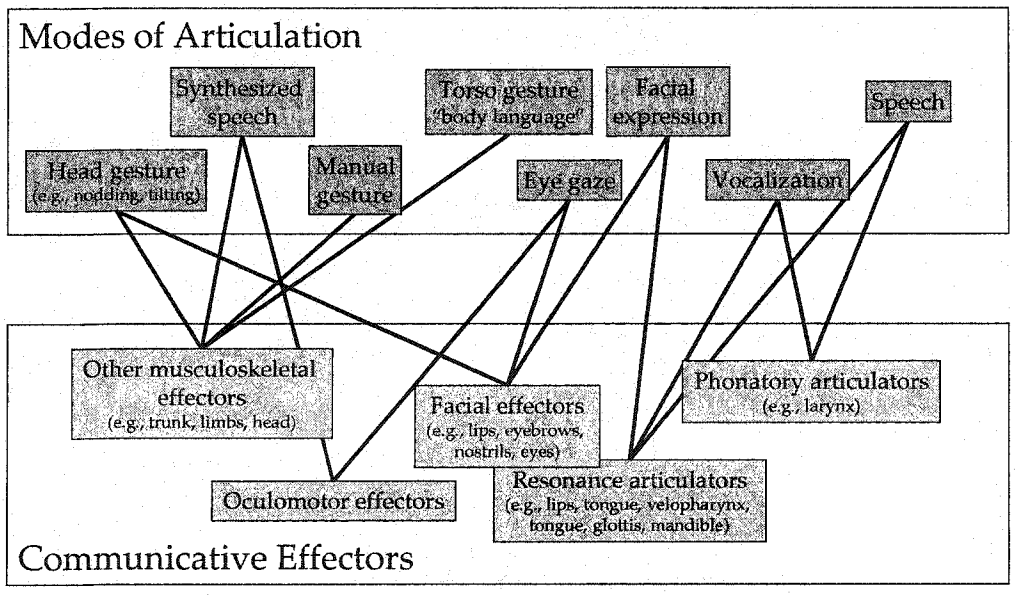
The aided mode of synthesized speech also requires the support of multiple underlying effectors. As described earlier in section 3.3.6, to operate the device, the user must perform some kind of input action — but whichever input actions are used, they require the support of effectors also underlie the use of other, unaided modes. Most VOCAs are accessed through input actions made using the hands; this requires the use of effectors that also support the mode of manual gesture. The use of the VOCA also requires the user to attend to the display; this requires the oculomotor effectors and the muscles of the head, so that the user can orient his or her face to the display, but these effectors that also support eye gaze and gestures of the head. If an eyebrow-activated input device is used, this requires the use of effectors that also support the mode of facial expression. When multiple modes require the support of one or more underlying communicative effectors, they can be seen as "competing" for a common resource. To provide the support required for the simultaneous use of multiple modes, an interlocutor must have sufficiently rapid, flexible, and coordinated motor control.

Not only do certain modes require the support of multiple effectors, but a single effector may provide the support needed by multiple modes. For instance, if an interlocutor produces facial expressions and spoken utterances simultaneously, then the lips, tongue, and mandible underlie the production of both sub-actions of speech and facial expression. In this case, multiple mode-specific sub-actions can be performed simultaneously (i.e., mode-specific sub-actions with respect to facial expression and the mode-specific sub-action with respect to speech do not conflict with one another).

Figure 4.1 An illustration of the dependencies of the modes on the underlying communicative effectors. Some modes are supported by multiple communicative effectors, and some communicative effectors support multiple modes. The bottom illustration shows a mode repertoire that includes synthesized speech.



(a) Mode repertoire of an unaided communicator.



(b) Mode repertoire of an aided communicator.

For an individual with a physical disorder, the requirements of multiple modes might not be simultaneously satisfiable. The disorder might cause motor movements to be slow, effortful, and poorly coordinated. Thus, constraints on the underlying physical effectors might affect not only the use of a mode, but also the degree to which multiple modes can be used *simultaneously*. When the support of the underlying effectors is inadequate for the simultaneous use of multiple modes, we describe those modes as *conflicting*. In particular, the design of the VOCA itself introduces mode conflict. As we have seen, an aided communicator may use his or her hands to provide input actions to the VOCA (in order to produce a "spoken" referring expression) or to produce a pointing gesture, but not both at the same time.

4.3 The Repertoire of Mode Strategies

4.3.1 Effect of a VOCA on a Communicator's Repertoire of Mode Strategies

As described previously, in section 3.3.3, individuals with communication disorders make use of a variety of communication strategies. In some situations, the communicator will make use of one or more unaided modes, whereas in others, he or she will use the aided mode of synthesized speech.

As described in section 4.2.4, the aided mode of synthesized speech conflicts with the unaided modes of gesture, facial expression, and eye gaze. The aided mode and these unaided modes rely on common underlying communicative effectors, which cannot provide adequate articulatory support. Thus, the use of the aided mode can come at the expense of the other modes of articulation. Conversely, the use of the unaided modes can preclude the use of the aided mode. If we think of an aided communicator's communicative effectors as resources, and each possible mode-specific sub-action as a potential consumer of those resources, then we can see that certain mode-specific sub-actions cannot be performed simultaneously due to inadequate articulatory support. However, mode-specific sub-actions that otherwise could have been performed simultaneously might instead be performed sequentially. An aided communicator, even with contraints on his or her communicative resources, can still make use of certain sequential multimodal strategies.

In many of situations in which the aided mode is used, the communicator modulates or augments the synthesized speech with communicative actions made by one or more of the unaided modes. For instance, once the aided communicator has produced VOCA input actions, he or she may produce gestures, facial expressions, and vocalizations to accompany the synthesized speech. This multimodal sequential strategy is an important one; with it the aided communicator can modulate the meaning of the synthesized speech with unaided mode-specific sub-actions (e.g., excited vocalizations or locative gestures). Such a strategy is synergistic because the overall communicative effect is derived from the modes working together (and is such that it cannot be derived from the use of either mode in isolation).

For the purpose of the discussion here, the three types of mode strategies that an aided communicator can potentially use will be labelled as follows:

- *unaided*: the mode strategies that involve the use of one or more unaided mode
- *aided-unimodal*: the mode strategy that involves the isolated use of the aided mode of synthe-

sized speech

- *joint aided-unaided*: mode strategies that involve the use of the aided mode in conjunction with the use of one or more unaided modes

When an individual uses a VOCA with a unimodal interface, its effect is to augment that person's repertoire of mode strategies from $\mathcal{R} = \{unaided\}$ to the repertoire $\mathcal{R}' = \{unaided, aided\text{-}unimodal, joint\ aided\text{-}unaided\}$. The AAC literature's consensus of the benefits of AAC devices for most individuals (see Beukelman and Mirenda [1998] for an summary) can be translated as meaning that, for most individuals, the repertoire $\mathcal{R}'$ is of better service than $\mathcal{R}$. The benefits of $\mathcal{R}'$ over $\mathcal{R}$ are especially apparent with unfamiliar communication partners.

4.3.2 Hypothesized Effect of Multimodal Interfaces to VOCAs

The bottleneck-reduction proposal, described in section 3.3.8, hypothesized that a multimodal AAC device will improve the effectiveness of the aided mode. But, as described in the previous section, the strategy of using the aided mode in isolation is merely one mode strategy of several. Will a VOCA with a multimodal interface improve the overall effectiveness of the aided communicator's repertoire of strategies over that with a unimodal interface?

With a VOCA with a multimodal interface (or multimodal VOCA, for short), the individual's repertoire of mode strategies will be augmented from $\mathcal{R} = \{unaided\}$ to the repertoire $\mathcal{R}'' = \{unaided, aided_{M}\text{-}unimodal, joint\ aided_{M}\text{-}unaided\}$, similar to the case with a unimodal VOCA. Shein et al. [1990] argued convincingly that the aided mode afforded by a multimodal VOCA will have many advantages over the aided mode that is afforded by a unimodal VOCA. Thus, we expect that the $aided_{M}$-unimodal mode strategy will have advantages over the aided-unimodal strategy that is afforded by a unimodal VOCA. But we observe that a multimodal AAC device will require multimodal input actions, which will recruit the use of even more of the aided communicator's resources. Essentially, the bottleneck reduction proposal is tantamount to a proposal to recruit more of the interlocutor's communicative resources for their interaction with the AAC device. But this "redeployment" of resources might mean that the aided communicator has fewer resources to support the use of the unaided modes. For example, the production of multimodal input actions may be so fatiguing that the aided communicator cannot, or is disinclined to, follow on with the use of an unaided mode. Therefore, it is quite possible that the collection of joint aided-unaided strategies that are afforded by a multimodal VOCA, as a whole, will be less effective than those afforded by a unimodal VOCA. The question for future VOCA design should be whether or not the benefits of a multimodal interface, in terms of the improvement to the aided-unimodal mode strategy, are sufficient to outweigh its disadvantages, in terms of detriment to the joint aided-unaided strategies.[2]

An additional issue is also relevant for future VOCA design. It is possible that communication partners will inaccurately perceive the $aided_{M}$-unimodal strategy as inherently "better" and discourage the aided communicator from using the other types of strategies, even if those strategies are actually more effective — Warrick [1988] described precisely this situation, which already occurs with unimodal VOCAs. With a multimodal VOCA, this bias toward the computer-supported mode might be exacerbated.

[2]This design issue was described previously — albeit using different terminology — in [Baljko, 2000b].

None of the follow-up work to date has acknowledged the problem that multimodal input actions may have a detrimental impact on the other, unaided modes of communication. This leads us to the statement of the problem that is the focus of this dissertation.

4.3.3 Statement of Problem

A thesis of this work is that, when comparing AAC devices, it is the effect of the devices on the interlocutor's *repertoire of strategies* that should be contrasted rather than just their effect on one specific mode strategy (such as the *aided-unimodal* strategy). The merits of repertoire $\mathcal{R}'$ over $\mathcal{R}$ are generally accepted, although they have not been formally evaluated. Is it the case that that the repertoire $\mathcal{R}''$ is better than $\mathcal{R}'$?

In order to compare AAC devices, we need to analyse the relative merits of different repertoires of strategies. A better understanding is needed of the relationship between, on the one hand, the individual's communicative resources and the interface of the VOCA and, on the other, the repertoire of mode strategies that they afford to that individual. The next chapters will describe a computational tool that illustrates these relationships.

4.4 Summary

In this chapter, we analyzed the *bottleneck reduction hypothesis* — that an AAC device with a multimodal interface might allow an individual to communicate more effectively than those that have unimodal interfaces. We developed a definition of a *repertoire of mode strategies* — an enumeration of the ways in which a communicator can make use of his or her various modes (such as vocalization, synthesized speech, facial expression, and gesture). We argued (1) that the bottleneck reduction hypothesis is tantamount to the recruitment of more of the aided communicator's communicative resources for the betterment of the particular strategy of using the mode of synthesized speech in isolation (the so-called *aided unimodal* strategy), and (2) that, through the mechanisms of *articulatory support* and *mode conflict*, improvements to this particular mode strategy may come at the detriment of the other mode strategies.

In this chapter, we formulated the central question of the present investigation: analysis shows that a multimodal VOCA will likely afford a more effective aided mode of synthesized speech than a unimodal VOCA, but is it necessarily the case that it will afford a better *repertoire of mode strategies* than a unimodal one? To answer this question, when comparing unimodal and multimodal VOCA interfaces, we must examine the *global* effect on a communicator's repertoire of mode strategies, as opposed to merely the *local* effect (i.e., the effect on the *aided-unimodal* strategy, which is one particular mode strategy in the individual's repertoire). A computational tool that demonstrates these two different types of effects, for each of the two interfaces, is described in the subsequent chapters.

Chapter 5

MSIM: The Computational Simulation of Mode Strategy Selection

5.1 Overview

To address the primary problem of this dissertation, we must investigate two interrelated mechanisms: the mechanism whereby a set of effectors and a VOCA *together* afford a repertoire of modes, and the mechanism whereby a repertoire of modes affords a repertoire of mode strategies. As this chapter will show, the repertoire of mode strategies that is available to a communicator is the product of a non-trivial number of factors that can interact with one another in subtle ways. It is important to have a computational simulation tool in order to demonstrate and to investigate these interrelated mechanisms.

This chapter will describe a simulation tool, called MSIM, that demonstrates computationally the relationship between, on the one hand, an individual's repertoire of mode strategies and, on the other, his or her communicative effectors and the interface of a VOCA. The previous chapter described the importance of this relationship: when considering the impact of a VOCA, we should consider the effect on *all* of the mode strategies in an individual's repertoire, and not solely the benefits with respect to the aided mode of synthesized speech, which is but one mode strategy of many. This is especially true when comparing unimodal and multimodal VOCAs. Both types afford the additional mode of synthesized speech, with the latter affording a more effective mode strategy of using synthesized speech in isolation. However, the latter also introduces increased mode conflict, the overall effect of which on the individual's repertoire of mode strategies might outweigh the benefits gained with respect to one particular strategy.

In section 5.2, an overview of the tool is given. The tool, in essence, instantiates two communicative agents, one which represents an aided communicator and the other which represents an unaided communicator, and places them in a situation in which they will become engaged in a specific joint activity. (These agents will be referred to as agent C (for *chooser*) and agent L (for *listener*),

for reasons to be described shortly.) The rationale for the focus on this specific joint activity, as well as the description of the joint activity itself, is provided in section 5.3. In essence, the joint activity restricts the types of behaviours that the architecture must implement. The architecture of the communicative agents is described in section 5.4. In MSIM, the aided communicator is considered to be a decision maker who, when choosing his or her multimodal communicative actions, considers the possible consequences. In section 5.5, the consequences considered by agent C are described — the aided communicator, when communicating, considers the satisfaction of the goal of being understood *and* of the goal to minimize physical effort, which is a challenge since these two goals often conflict. These consequences are relevant, since they reflect on agent C's evaluation of the mode strategies in its repertoire.

An additional issue is that an aided communicator's own effectors determine the types of multimodal communicative actions that can be performed in the first place. The tool MSIM is parameterized with respect to this and a number of other factors, which are described further in section 5.6. By specifying different parameter values, a unimodal VOCA, a multimodal VOCA, or no VOCA at all can be specified for agent C. The effectors and effector-mode interrelationships for agent C are also characterized by parameters of the tool. Last, in section 5.7, the use of MSIM is described again, this time in more specific detail.

5.2 Description of MSIM, the Simulation Tool

The simulation tool MSIM instantiates two communicative agents and places them in a situation in which they will engage in a specific joint activity (the architecture of the agents is tailored in order to bias them to perform this activity). The tool is presently configured so that one of the agents represents an aided communicator who uses a VOCA, and the other represents an unaided communicator.[1] The aided communicator is represented by agent C, and the unaided by agent L (C stands for *chooser* and L for *listener*, descriptors that will be explained in section 5.3).

MSIM provides a window on the inner workings of the architecture of agent C when it is contemplating the multimodal communicative action to perform. The architecture derives the repertoire of mode strategies that is available to agent C, on the basis of agent C's repertoire of modes, its VOCA, and its communicator-mode interrelationships (which are all specified as parameter values to MSIM). The repertoire of mode strategies includes both unaided and aided strategies. MSIM evaluates each of the mode strategies in its repertoire, given the current context, and chooses the best one for its multimodal communicative action.

The tool is not intended to be a predictive model of the multimodal communicative behaviour of aided communicators and has not been evaluated as such. The tool does not provide animations of the communicators, but rather represents their multimodal behaviour symbolically. The tool illustrates the hypothesized impact of a multimodal VOCA interface on the repertoire of mode strategies, via the mechanisms whereby a set of effectors and a VOCA *together* afford a repertoire of

[1] There is nothing in the design of MSIM that precludes other types of dyads or even giving the user the option of specifying the composition of the dyad (e.g., between two aided communicators or two unaided communicators). Polyads — exchanges that involve more than two interlocutors — could even be specified, although the agent architecture has not been developed to produce the behaviours specifically required for this scenario. Dyads other than the aided-unaided one described above and polyads were not the focus here and were not explored.

modes and whereby a repertoire of modes affords a repertoire of mode strategies. We accomplish this by contrasting the mode strategy repertoire that is afforded by a unimodal interface with the repertoire that is afforded by a multimodal interface. By developing MSIM, we provide a means to answer the question of whether the mode strategies that are afforded to an individual using a multimodal VOCA are better, as a whole, than the repertoire that is afforded to the same individual if he or she were to use a unimodal VOCA.

5.3 The Joint Activity

The simulation tool implements an approach whereby the joint activity is specified *a priori*. The agents, in their attempt to perform the joint activity, will derive communication goals, communication plans to satisfy these goals, and ultimately, communicative actions that implement the communication plans. In this thesis, the *multimodal referential communication task* has been chosen as the joint activity.[2] The focus on a specific joint activity allows us to make assumptions about the types of communication plans that the agents will need to produce, which allows us to abstract away the task of implementing a theory of action for the communicative agents. This particular task was selected for a number of reasons: it is an established technique for eliciting communicative actions; the production of referring actions is a basic communication skill; and it can be performed using a variety of different mode strategies (which include, to use the categories introduced in section 4.3, *unaided*, *aided-isolated*, and *joint aided-unaided* mode strategies).

The **referential communication task** has been used in psycholinguistic research to elicit utterances that convey definite reference. Krauss and Weinheimer [1964, 1966, 1967] used a task in which one subject must get the other to arrange ten hard-to-describe figures in a particular order; the subjects could not see each other but the referents were visible to both. In subsequent work, Clark [1992; 1996] used a similar type of collaborative task, although Tangram figures were used instead. At the time, these tasks were intended as instruments for exploring the role of accumulating common ground in the production of utterances that convey definite reference. But this task also serves to elicit a particular type of communicative action — definite referring expressions — and hence also a corresponding class of specific, underlying communicative plans.

Since the referential communication task, as it was originally specified, required that the subjects not be visible to one another, interlocutors tended to rely solely on the mode of speech. To elicit *multimodal* utterances, the task must be modified so that the interlocutors are inclined to draw on the various multimodal strategies. This is the **multimodal referential communication (MMRC) task**, which is performed as follows:

- Two participants face one another, with a set of objects positioned on a table between them.
- One subject, say C (for "chooser") selects a target object from the set and communicates its identity to the other subject, say L (for "listener"). Any desired modes, alone or in combination, can be used.
- L indicates his or her interpretation of the entity that C chose.

[2] Future versions of MSIM might allow the user to specify the particular type of joint activity.

Table 5.1 Steps required for the multimodal referential communication task.

Step	Agent C	Agent L
1	Identify target referent	wait
2	Derive communicative plan; pass to surface realization module; perform chosen surface realization	wait
3	wait	Attend to and interpret multimodal communicative action; indicate identity of interpreted intended referent.

For this joint activity, MSIM assumes that the aided communicator has been assigned to the role of C, the *chooser*, and the unaided communicator has been assigned to the role of L, the *listener*.

5.4 The Agents

MSIM instantiates both communicative agents from a common agent architecture, which has a plan derivation module and a surface realization module.[3]

The *plan derivation module* determines the type of action the agent should perform, if any. MSIM implements this in a very simple way. The agent is endowed with the knowledge that the task consists of three steps and the templates of the actions that are appropriate for each of the steps. The agent simply tracks the current state of the task and generates a plan that corresponds to the appropriate template. The steps of the task are given in table 5.1.[4] The agent's initiative to act is derived solely on the basis of the present state of the interaction, which means that the agent architecture makes use of a fixed-initiative mechanism, rather than a more sophisticated one based on mixed initiative.[5] Only one agent acts in each step of the activity. Following the performance of an action, the task progresses to the next step.

When the joint activity is in step 1, agent C identifies the target referent. The selection is presently made by random selection from among the set of potential referents in the agents' knowledge base (which is shared). This action is not manifested outwardly to agent L.

When the joint activity is in step 2, agent C's plan derivation module is responsible for generating a communication plan for the target referent. The plan is a functional specification of the action — i.e., *what* the action is intended to accomplish, but not *how* it will be accomplished. In MSIM, this step entails deriving a set of semantic primitives that serve to uniquely distinguish the target

[3] **Implementation Note:** The agent that represents the chooser and the agent that represents the listener are both instances of the `Agent` class, defined in the `MMSimulation` package. The parameters that are required by the `Agent` constructor are described in section 5.6.

[4] Moreover, the agents cannot misinterpret the present state of the joint activity, since the state transitions are unambiguous. For future work, this strong assumption should be relaxed; information about the current state might be imperfect, and the current state need not be always be determined correctly and with complete certainty (i.e., the state might be partially observable, rather than fully observable as it is in the present implementation).

[5] This fixed-initiative mechanism for the agent's high-level behaviours will likely not be feasible, nor desirable, in future versions of the simulation tool, which will seek to account for a larger set of joint activities. Ideally, each agent's behaviour would be driven by the agent's desire to fulfill its goals, and the simulation would be a mixed-initiative multi-agent system. In more sophisticated agent architectures, the agent would advance the satisfaction of its goals by first determining whether it should take the initiative or not, rather than having its initiative predetermined on the basis of the state.

from the competitor objects. This plan is subsequently passed to the multimodal surface realization module. These two modules implement the production process as it will be characterized in section 6.2.1.

The multimodal surface realization module in the agent architecture takes into account the characteristics of the agent's own communicative effectors, its VOCA, and the effector-mode interrelationships in order to determine a set of candidate surface realizations that are performable by the agent. It also takes into account the characteristics of the communication partner and the communicator's own sensitivity to fatigue when choosing from among the candidates. These characteristics relate to two different and often-conflicting types of goals that the agent is attempting to satisfy:

- the goal to be understood, and
- the goal to exert as little physical effort as possible.

These goals will be described further in section 5.5. **But we are most interested in the set of candidate multimodal surface realizations that the module derives and evaluates rather than in the specific one that is finally selected, since it is the candidate set that provides the basis for the analysis of agent C's repertoire of mode strategies.**

The architecture does not include any mechanism for memory or learning, which means that even if the agents have already performed the task a number of times, the agent's behaviour will not improve on the basis of past failures or successes.

In MSIM, the characteristics of agent C that relate to its ability to perform multimodal communicative actions, such as its repertoire of modes and whether it uses a VOCA or not, are parameters of the simulation. Also, the characteristics of agent L that relate to its ability to interpret multimodal communicative actions, such as its familiarity with AAC techniques, are parameters of the simulation. We use the term *simulation condition* to describe a particular configuration of these (and other) parameter values. MSIM expects its input file to contain a specification of these parameters for one or more simulation conditions (the parameters and their possible values will be described further in section 5.6).

Pseudo-code for the behaviour of agent C is given in table 5.2 and the multimodal surface realization module will be described in the subsequent chapter. Notice that in step 2b in table 5.2, there may be ties for the "best" candidate. In this case, the agent selects one of them randomly. In order to get a representative sample of the mode strategy employed by agent C, the same condition needs to be invoked a number of times. MSIM expects its input file to contain a specification of the number of times a particular simulation condition should be invoked.

When step 2 of the multimodal referential communication task is complete, agent L performs step 3 (attends to and interprets agent C's action).[6] Following this, MSIM records whether agent L's interpretation was indeed the correct one. Using this information, analyses can be performed to correlate agent C's mode strategy selection and the joint activity outcome. Once the agents have

[6] Agent L perceives agent C's action through its perceptual layer. Sensory-perceptual processing in the agent architecture is not a focus for this particular prototype and is implemented as though the agent has perfect information about the actions that the other agent has performed. The sensory mechanism is assumed to be free of disorder. Future modification to the agent architecture might include a more sophisticated mechanism for sensory-perception, so that sensory-perceptual disorders, in addition to articulatory disorders, might be simulated.

Table 5.2 Pseudo-code for agent C when performing step 2 of the multimodal referential communication task.

When in step 2 ...
2.1 Derive communicative plan
  2.1a Choose intended referent e_i
  2.1b Generate communication plan *plan-ref*(e_i)
2.2 Derive multimodal surface realization
  2.2a Generate Γ(*plan-ref*$(e_i)) = \{A_1, \ldots, A_n\}$, the set of candidate multimodal surface realizations.
  2.2b Choose "best" candidate $A_i \in \Gamma$
2.3 Perform multimodal communicative action with surface realization A_i

completed their task, the simulation driver suspends them. This completes a single invocation of a simulation condition. The input file specifies the number of times each simulation condition should be invoked by MSIM.

5.5 The Aided Communicator as a Decision Maker

A human or computational agent, when faced with the task of choosing from among a set of *alternatives* in light of their possible consequences, is characterized as a *decision maker*. A communicator, and an aided communicator in particular, can be regarded as a decision maker in at least two ways. First, throughout a communicative exchange, the aided communicator is faced with the decision of which communicative action to perform, if any. This is decision-making with respect to *plan derivation*. The decision maker's set of alternatives is the set of possible types of action plans that might be implemented. Second, given one of these communication plans, the aided communicator must decide which temporally-coordinated set of mode-specific sub-actions should be performed to realize that plan. This is decision-making with respect to *multimodal surface realization*. In MSIM, we implement decision-making with respect to multimodal surface realization in order to demonstrate the key ideas of this dissertation.[7] We make use of a specific task, which elicits only a specific type of communication plan, in order to abstract away the issue of how to best implement a theory of action for the communicative agents (and to avoid an implementation of decision-making with respect to plan derivation).

In the decision-theoretic paradigm, an action has exactly one consequence, but the consequence may have a number of different *attributes*. If so, then it is a *multi-attribute consequence*.[8] Our model of the relevant attributes of possible consequence states is based on Clark's [1996] model that interlocutors have three types of goals: domain, procedural, and interpersonal. In the MMRC task, agent C's *domain goal* is to be understood (i.e., the goal is that its intended referent be successfully identified). The *procedural goal* is to minimize the physical effort expended on the communicative

[7] The distinction between these two levels of decision making was described earlier in [Baljko, 2001a].

[8] In the literature on decision theory, a course of action is described as having *one* consequence, which has multiple attributes (as opposed to an action having several consequences). Within the context of decision theory, *attributes* are those features of a consequence that are taken into account in the evaluation of this consequence by the decision maker. One speaks, more precisely, about *value-relevant attributes*.

action. The *interpersonal goal* is to maintain contact in the communicative exchange and to not violate the rules of social engagement (e.g., to avoid actions that are likely to cause the communication partner to withdraw, such as imposing undue burden, with respect to time or cognitive effort). In MSIM, the agent architecture has been implemented so that the agent makes a choice in considering the consequences with respect to two different attributes — the degrees to which the domain and the procedural goals have been advanced. Satisfaction of the interpersonal goal is planned for future work.

In order to model the decision-making process, a mapping is postulated from the space of possible courses of action (the *action space*) into the space of consequences (the *consequence space*). The mapping must be presumed either to be a point-to-point mapping or a point-to-set mapping. With the former presumption, the *deterministic* case, a given course of action is modeled as having a given and certain consequence. With the latter presumption, the case of *risk* or *uncertainty*, it is understood there is a set of consequences, any one of which might follow from a given course of action. The deterministic case, described as *decision-making under certainty*, is not as appealing as the other models of decision making due to its simplifying assumptions,[9] but since these assumptions do not adversely affect the core functions of MSIM, it will be employed here.

The degrees to which each goal have been attained are considered to be attributes of a possible state. In MSIM, attainment with respect to each goal is characterized by a real value from the interval $[0, 1]$ (where 0 represents complete failure and 1 represents full goal attainment). Attainment with respect to the domain and procedural goals is denoted by s_D and s_P respectively. Our state space is therefore defined by $\mathcal{S} \subseteq \{\langle s_D, s_P \rangle \mid s_D, s_P \in [0, 1]\}$.[10]

Agent C has *preferences* for some states over others. *Preference* is an ordering of alternatives according to the agent's "likes" and "dislikes". For instance, following step 3 of the MMRC task, the state $\langle 1, 0.9 \rangle$ (i.e., agent C has been completely understood as intended and has exerted little physical effort in the process) is preferred over state $\langle 0.5, 0.2 \rangle$ (i.e., agent C has been partially understood and has exerted considerable effort to achieve this poor result — agent L has an ambiguous understanding of the intended referent). Care must be taken that the ordering of multi-attribute consequences is consistent and does not contain circularities (i.e., the ordering must be transitive, if $X \succ Y$ and $Y \succ Z$ then $X \succ Z$) [Turchin et al., 1991–2003]. A preference representation function under certainty is described as a a *value function* (and under uncertainty is a *utility function*) [Dyer and Sarin, 1979, p. 810]. Since MSIM makes use of decision making under certainty, agent C makes

[9] In *decision under uncertainty*, there are several possible mutually-exclusive consequences for each alternative. The decision-making procedure depends on whether the probability of occurrence of each consequence is known or not. If the probability distributions are known, then the choice among alternatives is equivalent to a choice among probability distributions. A simulation tool can make use of a model that specifies the conditional probability of each of the various possible consequence states, given a current state and a particular action. The decision maker's preferences for the consequences of an alternative are described by a *utility function* (the analog of a value function), which permits calculation of the *expected utility* of each alternative (the likelihood of each alternative's consequences multiplied by its utility value). In this scenario, the decision maker chooses the alternative with the highest expected utility. (Turchin et al. [1991–2003] describes this scenario, in which the probability distributions are known, as *decision under risk*.) If the probability distributions are not known, then the decision maker cannot make use of calculations of expected utility for the possible alternatives. Two approaches might be employed. In the first, an alternative criterion of choice is adapted from the broader context of game theory (e.g., the minimax rule, using which says to choose the alternative of which the worst possible consequence is better than all of the other alternatives' worst-case consequences). The second approach is to attempt to model the unknown probability distributions (e.g., by making use of expert assessments, or through analysis of previous decisions made in similar circumstances).

[10] The state space is not defined by the infinite set $\mathcal{S} = \{\langle s_D, s_P \rangle \mid s_D, s_P \in [0, 1]\}$ because, as will be shown in the next chapter, only a finite set of values are defined for each of s_D and s_P.

use of a value function in order to choose the "best" multimodal surface realization for a given communication plan (the function will described in the section 6.5.5 of the next chapter).

Recall from section 5.4 that completion of the MMRC task is modeled as a series of three discrete steps. Thus, the task can be characterized in terms of four states and three state transitions (a state transition is accomplished by each step of the task).

The initial state and the state that follows the completion of step 1 are both assumed to be $\langle 0, 1 \rangle$. That is, $s_D = 0$ because the domain goal has not yet been satisfied (no interpretation has been made yet), and $s_P = 1$ because the procedural goal is fully satisfied (no physical effort has yet been expended). The state that follows step 2 is assumed to be of the form $\langle 0, s_P \rangle$, where the value s_P depends on the particular multimodal surface realization that has been selected and performed by agent C. We use the variable S_2 to represent whichever state follows the completion of step 2. The state that follows step 3 is assumed to be of the form $\langle s_D, s_P \rangle$, where the value of s_P is assumed to be unchanged from the previous state and the value of s_D depends on agent L's interpretation of the multimodal communicative action that has been performed by agent C. We use the variable S_3 to represent whichever state follows the completion of step 3. The architecture of agent C ensures that any communicative action implements the underlying communication plan, although the possible communicative actions vary with respect to the ease with which they might be interpreted. Therefore, the state transition model derives s_D on the basis of the degree to which agent L is familiar with agent C and the AAC system. (Modeling the impact of other factors on agent L's interpretive abilities is planned for future research.)

In MSIM, when agent C evaluates the possible multimodal surface realizations for its communicative action in step 1, it considers the *combined value* of the attributes s_D and s_P that the state transition model derives for state S_3. The model of the state transitions and the multi-attribute value function will be described in section 6.5.5.

5.6 Input to MSIM

The user prepares a specially-formatted text file containing the specification of one or more simulation conditions, and passes it to MSIM as an input.[11] As described previously in section 5.4, a *simulation condition* refers to a particular configuration of parameter values. Each specified simulation condition is invoked one or more times. Each simulation condition will be denoted by a label of the form $cond_i$, where i is the index of the condition, $1, \ldots, n$. An example file is provided in appendix A.4.

The specification of a simulation condition $cond_i$ entails specifying the value k_i, the the number of times the simulation condition should be invoked. Because the selection of mode strategy is stochastic, different invocations might give different results.

The specification of a simulation condition $cond_i$ also entails specifying the set of entities that is arrayed between agent C and agent L in the MMRC task. Agent C might choose any entity from this set to be its intended referent. For each entity, a set of semantic primitives that uniquely

[11] **Implementation Note:** Presently, a text editor is used to prepare these input files manually. A application, called `ScenarioEditor`, is currently under development to provide a front-end to the input file. This application, which has a graphical user interface, will assist the user in preparing the input files to MSIM. This application will be more user friendly than a text editor, and will implement a verification mechanism to ensure that the input file is correctly formatted.

distinguish it from the other entities must also be specified. (This part of the specification has another component, but its description requires the use of definitions that are not yet provided; it will be introduced in section 6.3.)

The specification of a simulation condition $cond_i$ also entails specifying the characteristics of agent C's communicative effectors and effector-mode interrelationships.[12] This specification includes the following:

1. The set $\mathcal{M} = \{m_1, \ldots, m_n\}$, the participant's repertoire of modes of articulation. This set may include the mode of synthesized speech.

2. Additional functions R_1, R_2, C_1, and C_2 must be specified for the mode repertoire; these will be described in section 6.5.4.

3. The specification of a simulation condition $cond_i$ also entails specifying agent C's and agent L's *level of familiarity* with one another. The familiarity is given as a value from $[low, med, high]$. Agent C's *tolerance to fatigue* must also be specified, given as a value from $[low, med, high]$. These parameters characterize the communicative scenario, and are used by the architecture of agent C in order to tailor the evaluation of the candidate multimodal surface realizations (the effect of these parameter values will be described further in section 6.5.5).

4. The **support function** function $\mathcal{S}$. The requirements of each mode of articulation in terms of the support required from underlying effectors. $\mathcal{S}$ is characterized by:

$$\mathcal{S} : \mathcal{M} \longrightarrow \mathcal{P}(\mathcal{E}) \tag{5.1}$$

where the value $\mathcal{S}(M_i)$ is the set of communicative effectors that provide the support for the use of mode M_i. For example, the value $\mathcal{S}(M_i) = \{E_j, E_k\}$ means that the use of mode M_i requires the support of effectors E_j and E_k. The domain of the function is the mode set $\mathcal{M}$, and the range is $\mathcal{P}(\mathcal{E})$, the power set of $\mathcal{E}$ (the set of all subsets of $\mathcal{E}$):

$$\mathcal{P}(\mathcal{E}) = \{\{\}, \{E_1\}, \ldots, \{E_k\}, \ldots, \{E_1, E_2\}, \ldots, \{E_1, E_2, \ldots, E_k\}\}.$$

The power set $\mathcal{P}(\mathcal{E})$ enumerates all of the combinations in which the effectors might potentially be used. Typical effector-mode relationships are given in table 5.3.

The effector-mode interdependencies expressed by the support function $\mathcal{S}$ can alternatively be expressed by the mode repertoire's **interference set** I. I is a set of *mode-specific interference sets*, that is:

$$I = \{I_{M_1}, I_{M_2}, \ldots, I_{M_n}\} \subset \mathcal{P}(\mathcal{M}) \tag{5.2}$$

The interference set is a set of *mode-specific interference sets*. A mode-specific interference set I_{M_k} is the set of all of the modes whose simultaneous use with the mode M_k is not possible. $I_{M_k} = \emptyset$ indicates that no modes interfere with M_k, whereas $I_{M_k} = \mathcal{M} - M_k$ indicates that

[12]The characteristics of agent L's communicative effectors and effector-mode interrelationships must also be specified, since these values are needed in order to instantiate the agent, but since agent L does not produce multimodal communicative actions in the present version of MSIM, we will use simple default values and not consider this parameter further.

all the other modes interfere with M_k. Interference is symmetric, meaning that if mode M_j interferes with M_k, then $M_j \in I_{M_k}$ and $M_k \in I_{M_j}$.

Of particular interest are the interference sets of the aided mode that is afforded by a VOCA that has a unimodal interface, which is given in (5.3). This interference set show that use of the aided mode of synthesized speech conflicts with the use of the unaided modes of gesture and eye gaze.

$$I_{\text{U-VOCA}} = \{\text{MANUALGESTURE}, \text{EYEGAZE}, \text{HEADGESTURE}\} \tag{5.3}$$

Table 5.3 The effector-mode relationships for various modes of articulation, given in terms of values of the support function $\mathcal{S}$. U-VOCA refers to the mode of synthesized speech that is made possible through the use of a voice-output communication aid that has a unimodal interface. The modes SPEECH and VOCALIZATION are similar in that they both make use of the speech-sound articulators (vocalization is the articulation of non-speech sounds).

$$\begin{aligned}
\mathcal{S}(\text{MANUALGESTURE}) &= \{\text{HANDS}\} \\
\mathcal{S}(\text{HEADGESTURE}) &= \{\text{HEAD}\} \\
\mathcal{S}(\text{TORSOSHIFTS}) &= \{\text{TORSO}\} \\
\mathcal{S}(\text{FACIALEXP}) &= \{\text{EYES/EYEBROWS}, \text{NOSE}, \text{MOUTH}, \text{MANDIBLE}, \text{OCULOMOTOR}\} \\
\mathcal{S}(\text{EYEGAZE}) &= \{\text{OCULOMOTOR}\} \\
\mathcal{S}(\text{SPEECH}) &= \{\text{PHONATORY}, \text{RESONANCE}, \text{MANDIBLE}\} \\
\mathcal{S}(\text{VOCALIZATION}) &= \{\text{PHONATORY}, \text{RESONANCE}, \text{MANDIBLE}\} \\
\mathcal{S}(\text{U-VOCA}) &= \{\text{HANDS}, \text{OCULOMOTOR}, \text{HEAD}\}
\end{aligned}$$

5.7 Use of MSIM

During the invocation of each $cond_i$, agent C produces a multimodal communicative action in step 2 of the multimodal referential communication task. In doing so, the agent's architecture generates a set of candidate multimodal surface realizations and selects one from that set. Both the set of candidates and the selected candidate are considered to be the "outputs" of the simulation tool.[13] The candidate sets from different simulation conditions then can be compared and contrasted to one another. Differences in the characteristics of the candidate sets can be attributed to differences between the conditions under which the simulation of the joint activity was invoked. For instance, if a unimodal strategy becomes more highly rated in the candidate set from $cond_i$ than in the candidate set derived from $cond_j$, this difference can be attributed to the differences in the two conditions. The computational simulations are needed because the derived candidate sets are products of a non-trivial number of constraints that can interact with one another in subtle ways (the derivation

[13]The agent architecture evaluates of each of the candidates and the selection of the surface realization from among the candidates follows from their evaluations. Subsequent discussions will focus on the candidate sets (the candidate that was selected can be inferred from its evaluation).

of the candidate set will be described in section 6.3). The differences that are of most interest here are to be found among simulation conditions in which the type of VOCA that agent C (and the effector-mode interrelationships) are manipulated.

Thus, MSIM is intended to be used to gather information about the candidate sets that are generated in various conditions where those sets are a function of the simulation tool's parameter values. This interdependency is described further in the next chapter.

We hypothesize that this approach will also be useful for future simulations of joint activities other than the multimodal referential communication task. Many types of joint activities have designated roles for the interlocutors and involve predictable routines of communicative actions. These actions can be aligned with one another over simulation conditions, so that they may be compared and contrasted across different simulation conditions. In the present implementation, the underlying joint activity is highly structured, which makes it straightforward to identify the multimodal communicative actions that are alignable. In future versions of the simulation tool, additional techniques might be needed to identify the multimodal actions that should be aligned with different types of joint activities.

5.8 Summary

This chapter described the simulation tool MSIM, which provides a computational demonstration of the repertoire of mode strategies that is available to an aided communicator, who is represented in the simulations by a communicative agent denoted by C. MSIM implements the mechanism whereby a set of communicative effectors and a VOCA, together, afford a repertoire of mode strategies. The characteristics of the communicator's effectors and of the VOCA's interface are supplied to MSIM as parameter values.

MSIM focuses on a specific communicative activity — the multimodal referential communication task. The agent architecture exploits the assumption about the type of joint activity that is to be performed and implements only the specific behaviours that are required in order to engage in this task. This allows us to abstract away the task of communication plan generation and, instead, to focus on the derivation of multimodal surface realizations. In MSIM, the multimodal surface realizations that might possibly be selected by agent C (and, correspondingly, the mode strategies that are available for use) depend on its communicative effectors and the relationships between those effectors and the modes of articulation (which may include the aided mode of synthesized speech). The agent architecture implements decision making under certainty, which necessitated formalizing the performance of the MMRC task as a set of states and transitions between them. Agent C's selection of a particular multimodal surface realization accomplishes one state transistion.

In the next chapter, the derivation and evaluation of the set of candidate multimodal surface realizations, and the categorization of agent C's repertoire of mode strategies, are described in more detail.

Chapter 6

The Mode Strategy Selection Module

6.1 Overview

Previously, we characterized the aided communicator as a decision maker who, given a communicative plan, must choose a temporally-coordinated set of mode-specific sub-actions to realize that plan, thereby selecting an overall mode strategy for the communicative plan. This chapter describes the module of the agent architecture responsible for deriving a multimodal surface realization for a communication plan. The module performs two operations: (1) the generation of a set of candidate surface realizations; and (2) the evaluation and selection of the "best" surface realization from among the candidates.

The candidate generation process itself is based on an algorithm that derives a set of matrices, each of which represents a multimodal surface realization. Each of the candidate surface realizations satisfies criteria which stipulate that the surface realization must be performable by the agent (given its available communicative resources) and that the underlying communication plan must be properly realized. The software implementation of this algorithm is described here along with some example outputs. Further techniques will be described for determining the mode strategy associated with a specific candidate the surface realization and for categorizing the repertoire of mode strategies corresponding to a set of candidates. In section 6.5, the evaluation of the candidates will be described. This evaluation is done by the agent architecture in the context of its decision-making process — in order for the agent to perform its task, it must identify the best candidate. But the evaluation also serves another purpose — the candidates are categorized according to mode strategy used and the evaluations of the candidates that correspond to each mode strategy, collectively, are used to characterize the mode strategy. In this way, each of the mode strategies in the agent's repertoire can be characterized. This analysis technique will be used in chapter 7, where we will provide a solution to the central problem of this thesis, which is to illustrate the impact of a multimodal VOCA on an interlocutor's repertoire of modes and contrast it to that of a unimodal VOCA.

6.2 Architecture of the Multimodal Surface Realization Module

6.2.1 Characterization of the Production Process

The process whereby multimodal communicative actions are produced by human communicators has been characterized in a number of different ways: as a process through which communicative functions are *realized* by a set of behaviours [Cassell et al., 2000, p. 35]; as a process in which chunks of information are *produced* [Martin et al., 2001, p. 2]; as a process in which communicative actions are *constructed* to achieve given goals [Cassell and Stone, 1999, p. 38]; and as a process in which messages are *conveyed* by one or more modes [Martin and Béroule, 1995, p. 24]. In some way, all of these characterizations conceptualize the production of communicative actions as a process through which some sort of an underlying, mental entity is derived and then gets "realized" by an observable instance of behaviour.[1] In MSIM, the process whereby an agent produces a multimodal communicative action takes place in two stages: plan derivation and surface realization.[2]

6.2.2 Plan Derivation

The mental entity that is derived in the first stage of the production process has been variously characterized in terms of communicative plans, communicative functions, chunks of information, goals, or message content. Although the precise nature of this mental entity is still under discussion, it is generally agreed that it does exist, in some form.

Agents in MSIM construct a *communication plan* during the plan derivation stage of the production process. The derivation of the communication plan is the derivation of *what* a communicative action is intended to accomplish (i.e., its function or purpose). It does not also include the specification of *how* the action should be performed.

The process whereby communicative plans are derived has been modeled in several different ways. The first (and most common) type of model makes use of the AI planning paradigm. Heeman and Hirst [1995] and Traum and Allen [1994], in particular, formulated it as part of a collaborative process. Another type of model makes use of probabilistic decision-making mechanisms. In the *Quartet* architecture, developed by Paek and Horvitz (see [Horvitz and Paek, 1999, 2000; Paek and Horvitz, 1999, 2000]), communicative plans are an intermediate level of representation, which is seen as characterizing the "grounding strategy" that the agent is to adopt. In this process, the communication plan is derived iteratively, during which its semantic content is successively refined through a probabilistic mechanism called value-of-information analyses.[3] The rule-based approach of Poggi and Pelachaud [1998] and Cassell et al. [2000] made use of predefined templates, such as those given in examples (1) and (2) below.

[1] The *output* of the process of multimodal communicative action production must be distinguished from the process itself. McNeill [1992] distinguishes between the output or *surface realization* of an utterance and the *utterance* itself — "an utterance is ... a process that has an internal development and has ... surface linguistic constituents [in] its final stage" [p. 218]. We, too, conceive of communicative actions as consisting of "internal development" and of a final component in which surface-level behaviours are produced. However, in this work, what McNeill describes as "linguistic constituents" are assumed to be any mode-specific sub-actions (and not just speech-specific sub-actions, such as the articulation of words).

[2] The module described here is based on the one described in [Baljko, 2000a].

[3] This refinement process is only applied to a subset of communicative actions — those that concern the agent's acknowledgment of uncertainty, as opposed to utterances that would carry out actions.

(1) $\mathcal{S}$ INFORM $\mathcal{L}$ THAT $\mathcal{X}$
[Poggi and Pelachaud, 1998]

(2) SPEECH ACT TEMPLATE: Describe(object $\mathcal{Y}$, aspect $\mathcal{Z}$)
[Cassell et al., 2000][4]

Recall that agent C, in step 1 of the MMRC task, chooses a referent $e_i \in E = \{e_1, \ldots, e_q\}$ and designs a multimodal communicative action to identify it. The first stage of designing the communicative action is for agent C to derive a communication plan. In MSIM, the assumption has been made that this communication plan always will consist of:

- a set of **semantic primitives** $X_i = \{p_1, p_2, \ldots, p_{n_i}\}$. Each primitive $p_j \in X_i$ serves to provide information that discriminates the intended referent from the set of potential referents. The set $\mathcal{X}$ is the set of all of the semantic primitives that are known to an interlocutor.
- a partial ordering $\mathcal{O}_i = \{(p_j, p_k) \mid p_j < p_k, p_j, p_k \in X_i\}$ on the set X_i (which is transitive, but non-reflexive and antisymmetric). The ordering relation is a generalization of the surface ordering that is observed to hold for linguistic modes (e.g., *the big red ball* but not $*$*the red big ball*).

The sets $X = \{X_1, \ldots, X_q\}$ and $\mathcal{O} = \{\mathcal{O}_1, \ldots, \mathcal{O}_q\}$ for each potential referent are passed to MSIM as parameters. All communication plans of this form will be denoted by *plan-ref*(e_i).

6.2.3 Surface Realization

Once the communication plan has been derived, a surface realization is derived for it. In most multimodal agent architectures and applications, communication plans (or functional specifications of actions) typically stand in a one-to-one relationship with surface-level behaviours (even though those behaviours might be only one of several ways in which the plan might be realized by human communicators). For instance, in the *Quartet* architecture [Paek and Horvitz, 2000], the agent's on-screen embodiment includes communicative effectors for facial expression and arm, torso, and body gesture. This repertoire of modes affords various multimodal strategies, but the process of surface realization simply made use of predefined templates.

In the FMBT architecture for the embodied communicative agent REA [Cassell et al., 2000], surface realization is accomplished by the natural language generation engine SPUD [Stone, 2001], which was extended to produce multimodal surface realizations. But the generation engine does not tailor the surface realization to suit the communicative context or communication partner in any particular way. The multimodal incremental algorithm developed by van der Sluis and Krahmer [2000] (also see [van der Sluis, 2001]) did provide a means by which the surface realization might be tailored, but the tailoring process was based solely on the factor of the interlocutor's proximity to the intended referent. Factors that relate to the availability or condition of the interlocutor's repertoire of modes were not explicitly modeled. The mechanism that is presented in this thesis differs

[4] In the work of Cassell et al. [2000, p. 35], these constructs are called *communicative functions* rather than communicative plans.

from that of van der Sluis and Krahmer because the resources with which the surface realization will be performed are explicitly represented.

In human communicative processes, it is often the case that any one of multiple possible surface realizations may serve to implement a given plan. In previous research on models of surface realization for referring actions, the surface actions that compose these behaviours are typically construed to be surface *speech* actions — e.g., "surface speech actions correspond to the components of the description" [Heeman and Hirst, 1995, p. 355] and "a set of attribute-value pairs ... distinguish an entity [e.g., the intended referent] from a set of entities [e.g., the other potential referents]" and these attributes "are realizable by *absolute* adjectives" [Dale, 1989, pp. 71–72]. This technique can be seen as implementing the generation of *unimodal* surface realizations for referring actions. In these approaches, the intended referent is represented by a set of semantic primitives, each of which must be signalled by a speech-specific action (e.g., adjectives, or lexemes more generally).

Unlike unimodal surface realization, in which each semantic primitive in a communication plan gets realized by one speech-specific action, multimodal surface realization allows different types of mode-specific sub-actions to realize the semantic primitives in a communication plan. In particular, the principle of *switching* or *substitution* might be used. This principle was first described by Kendon [1988], who characterized gestures and facial expressions as "replacements" for a single word or a single sentence component (such as complex descriptive phrases). Example (3) below [Kendon, 1988, p. 135] illustrates a situation in which a semantic primitive is being signalled by both facial and gestural sub-actions. Example (4) is a variation in which a different facial expression and a different gesture are used. The semantic primitive might otherwise have been realized using a lexicalized description, such as in example (5). (Sub-actions that are in angle-brackets and co-indexed are performed simultaneously.) Example (5) also shows the situation in which the semantic-primitives are realized using *only* the mode of speech.

(3)
B: *Speech:* Their parents are professors but the kids are $\langle\ \rangle_1$.
Facial Exp: ⟨"disgusted" facial expression⟩$_1$
Gesture: ⟨rapidly moving both hands forward, splaying out her fingers to the fullest⟩$_1$

(4)
B: *Speech:* Their parents are professors but the kids are $\langle\ \rangle_1$.
Facial Exp: ⟨tongue sticks out⟩$_1$
Gesture: ⟨"thumbs down" gesture⟩$_1$

(5)
B: *Speech:* Their parents are professors but the kids are really detestable.
Facial Exp: (NOTE: not used)
Gesture: (NOTE: not used)

This situation, in which a communicator uses certain modes for some components of the semantic content of a communicative action and then changes to other modes for other components, has been described as *mode switching* [Martin and Béroule, 1995]. It has been hypothesized that it is motivated by mode *complementarity*, the notion that the modes in a repertoire each have semantic-primitive-specific advantages and disadvantages. Conversely, the existence of an *equivalence* relation among modes (and in certain situations) has been asserted [Martin et al., 2001; Kipp, 2001]. (This assertion entails that an interlocutor can switch from the use of one mode to another without

disadvantage, or conversely, without additional benefit.) Presumably, whether multiple mode-specific sub-actions are equivalent or not also depends on the communication partner, in addition to the particular semantic primitive to be realized. Further clarification of this issue and of which modes are complementary and which modes are equivalent has not been provided in the research literature. Any such account, however, must acknowledge that the advantages and disadvantages of a mode for realizing a given semantic primitive is not only a function of the particular semantic primitive, but also a function of the communication partner.

An idea that has intuitive appeal is that the use of multiple modes creates redundancy, which can then be subsequently exploited by the addressee (e.g., by forming an enriched percept, which has a better chance of being interpreted as intended). This idea has been described as *mode redundancy* [Martin and Béroule, 1995]. We define a *redundant* mode strategy as one in which at least one semantic primitive is realized by multiple mode-specific sub-actions. Different degrees of redundancy can be distinguished on the basis of the number of semantic primitives that are realized by multiple mode-specific sub-actions. When a semantic primitive is signalled by multiple mode-specific sub-actions, we describe it as *multiply signalled*.

In MSIM, we assume that each semantic primitive may realized by speech sub-actions or sub-actions performed with other modes. For example, iconic gestures can serve to signal semantic primitives that are concrete nouns (see section 4.2.2). Deictic gestures (e.g., pointing gestures using the hands and directed, prolonged gaze) can also signal semantic primitives by their locative action. More generally, we hypothesize that any one of a *set* of sub-actions can serve to signal semantic primitive p_j, and that communication partners can successfully interpret a particular primitive p_j provided that **at least one** of the sub-actions from this set has been performed to signal it (and the mode need not be speech).

In MSIM, the derivation of a multimodal surface realization for a given communication plan is accomplished in two stages: first, a set of candidate surface realizations is generated, and, next, the "best" candidate is selected from among them. Pseudo-code for this component of agent C's behaviour is given in table 6.1 (an elaborated version of table 5.2). This approach entails the generation of many surface realizations, only one of which eventually gets used. Many of the candidates might not even be serious contenders. This approach has the disadvantage of computational excess; on the other hand, it is modular and conceptually straightforward, which is a more important consideration at this early stage of investigation.

A surface realization is defined to be a set of temporally-coordinated mode-specific sub-actions. Surface realizations will be denoted by A_i, where the index i serves to distinguish among different surface realizations. An essential property of the agent architecture used in MSIM is that a given communication plan may be realized by any one of many possible surface realizations.[5] Thus, the process of multimodal surface realization in MSIM can be conceptualized as

[5]Speech Act Theory makes an analogous distinction for spoken acts of communication: the theory tells us that for each physical utterance, three acts are actually performed Austin [1962] — a locutionary act (the act of uttering a sequence of words, such as shouting or whispering), an illocutionary act (the act performed *in* saying, such as requesting, asking, telling, suggesting, or greeting), and a perlocutionary act (the act that is the actual result of the utterance, such as impressing, persuading, or embarrassing). Locutionary acts and illocutionary acts stand in a many-to-many relationship with one another; multiple possible illocutionary acts can correspond to a particular locutionary act, and a particular illocutionary act can be accomplished by multiple possible locutionary acts. If multimodal actions are considered, then there are even more possible locutionary acts that could accomplish a particular illocutionary act. For example, to refer to an entity, a speaker might use, in isolation or in combination, the modes of speech (e.g., through the use of various deictic linguistic

Table 6.1 Pseudo-code for agent $\mathcal{C}$ when performing step 2 of the multimodal referential communication task (an elaborated version of table 5.2).

When in step 2 ...
2.1 Derive communicative plan
  2.1a Choose intended referent e_i
  2.1b Generate communication plan *plan-ref*(e_i)
2.2 Derive multimodal surface realization
  2.2a$_1$ Generate Γ(*plan-ref*(e_i)) $= \{A_1, \ldots, A_n\}$, the set of candidate multimodal surface realizations.
  2.2a$_2$ Choose value function $V_j \in V$.
  2.2a$_3$ Calculate the *value* of each candidate $V_j(A_i)$ $\forall i = 1, \ldots, n$
  2.2a$_4$ Derive $\Gamma' \subseteq \Gamma$, where $\Gamma' = \{A_i \in \Gamma \mid |\mathtt{max}\{V_j(A_1), \ldots, V_j(A_n)\} - V_j(A_i)| \leq \delta\}$
  2.2b Choose "best" candidate $A_i \in \Gamma$: select A_i randomly from Γ'
2.3 Perform multimodal communicative action that has surface realization A_i

the construction of a mapping between a given communication plan and a set of surface realizations for it. This mapping can be alternatively expressed as a set of ordered pairs of the form $\{$(*plan-ref*$(e_i), A_1), \ldots,$ (*plan-ref*$(e_i), A_n)\}$, where $A_1, \ldots, A_n$ are the candidate surface realizations for the communication plan *plan-ref*(e_i). We hypothesize that, once the communication plan is established, the interlocutor's task can be characterized as choosing from among a set of candidate surface realizations $A_1, \ldots, A_n$, and that, in choosing a surface realization, a communicator implicitly chooses a mode strategy.

6.2.4 The Representation of Surface Realizations Using Matrices

In MSIM, candidate surface realizations are represented using *surface realization matrices*. The rows of a surface realization matrix correspond to the communicator's mode repertoire. If a communicator's mode repertoire is $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$, then the number of rows in A_i is n. The columns of the matrix represent discrete timesteps of the communicative action. The timestep granularity — termed δ_t here — must be chosen in advance. The value $\delta_t = 33.\dot{3}$ msec is a convenient value; with it, each column corresponds to one frame of video, assuming a frame rate of 30 frames/second. This value is useful for later validation purposes involving video analysis of real human communication. Larger values of δ_t would correspond to a coarser level of representation. The number of columns in A_i is given by τ, an upper bound on the number of timesteps.

The elements of the matrix correspond to labels of mode-specific sub-actions. To derive these labels, we must define inventories of sub-actions for each mode and label each one with an index value. The labelling must be done carefully; since new sub-actions might be added to the inventories. We assume only that the sub-actions in an inventory can be associated with unique identifiers, such as non-negative integers (this assumes only that the mode inventories are at most countably infinite).

expressions), gesture (e.g., through the use of various types of gestures, deictic or otherwise), facial expression, gaze, torso and head movement and so on.

The elements of the surface realization matrix A_i are defined as follows:

$$A_i[j,k] = \begin{cases} 0 & \text{no sub-action is articulated using mode } j \text{ at time step } k, \\ l & \text{the sub-action with index } l \text{ is being performed using mode } j \text{ at time step } k \end{cases} \quad (6.1)$$

The representation of a multimodal communicative action using a surface realization matrix provides a detailed representation of multimodal communicative actions (i.e., in terms of which modes of communication were used, how, and when). Please refer to appendix A.3 for a description of the previous research upon which this formalism is based.

Provided that a sub-action is performed by one and only one mode and that a mode can be used to perform only one sub-action at a time, any multimodal communicative action that can be represented using the timeline-based formalism can also be represented using the matrix-based formalism. Specifically, the timeline-based representation will contain no intervals that overlap with one another within a single row (which would cause a problem for a matrix, since each element in a row can only represent a single action). In addition, there will be no sub-actions that require intervals that span multiple rows.

6.3 The Generation of Candidate Surface Realizations

In this section, we describe a technique for deriving Γ(*plan-ref*(e_i)), the set of multimodal communicative surface realizations that both realize the communication plan *plan-ref*(e_i) *and* can be performed by agent C.

This technique requires the definition of a set of zero or more mode-specific sub-actions that realize each of the semantic primitives that might possibly occur in a communication plan that is constructed by agent C. We denote this set as follows:

$$\mathcal{R}(p_i) = \{a \mid a \text{ is a mode-specific sub-action that serves to realize the semantic primitive } p_i\} \quad (6.2)$$

This set must be specified as a parameter value when agent C is constructed.[6]

This technique also requires the definition of the mode-specific sub-actions that can be performed by agent C. We define these sets as follows:

$$\mathcal{A}(M_j) = \{a_{j,1}, \ldots, a_{j,n_j} \mid a_{j,k} \text{ is a sub-action that can be performed by agent } C \text{ using mode } M_j\} \quad (6.3)$$

Obviously, for agent C to have any hope of implementing a communication plan, the intersection of $\mathcal{R}(p_i)$ and $\{\mathcal{A}(M_1), \ldots, \mathcal{A}(M_n)\}$ must not be empty (for each of the primitives p_i in the communication plan).

For the purpose of the Δ criterion below, a set of triples must be defined to represent the minimum and maximum durations of time (given in timesteps) in which the sub-action $a_{j,l}$ can be performed:

$$\Delta_{M_j} = \{a_{j,l}, \delta_{\text{j,l,min}}, \delta_{\text{j,l,max}} \mid \forall\, a_{j,l} \in \mathcal{A}(M_j)\} \quad (6.4)$$

[6]Recall from section 6.2.2 that the set $\mathcal{X}$ is the set of all of the semantic primitives that are known to the agent. In section 7.2, the sets $\mathcal{R}(p_i)$, $\forall p_i \in \mathcal{X}$ are specified for four simulation conditions.

For the purpose of the completeness criterion below, we define the set of mode-specific sub-actions that both signal p_i *and* that can be performed by a particular interlocutor by:

$$\mathcal{A}_{|p_i} = \{\mathcal{R}(p_i) \cap \mathcal{A}(M_1), \ldots, \mathcal{R}(p_i) \cap \mathcal{A}(M_n)\} \tag{6.5}$$

An advantage of the matrix-based representation formalism is that we can begin by considering the set of **all** $m \times t$ matrices of non-negative integers, for all $m, t > 0$. We regard this set as representing the set of all possible surface realizations, although many of these would correspond to nonsensical communicative actions if they were to be performed by a human communicator. We next define additional criteria (the *mode inventory*, the *mode conflict avoidance*, and the Δ criterion) in order to filter out the surface realizations that cannot be articulated by agent C (e.g., due to constraints on its mode repertoire or mode conflict). Next, we define additional criteria (the *completeness* and the *preservation of ordering* criteria) in order to identify which among the articulable surface realizations actually serve to implement the given communication plan, *plan-ref*(e_i). The set of matrices which satisfy all of the above criteria is Γ(*plan-ref*(e_i)) (or Γ for brevity).

6.3.1 The Mode Inventory Criterion

The **mode inventory criterion** stipulates that the values of the matrix elements must correspond to valid mode-specific sub-actions. The i-th row of a surface realization matrix represents the actions that are articulated by the mode M_i, and the integer values that are valid for the elements of a particular row of a matrix must be derived from the associated inventory of mode-specific sub-actions. In other words, every matrix $A_i \in \Gamma$(*plan-ref*(e_i)) must meet the definition of a surface realization matrix that was given in (6.1) in section 6.2.4.

6.3.2 The Mode Conflict Avoidance Criterion

The **mode conflict avoidance criterion** stipulates that the communicative agent must have adequate articulatory support for all of the mode-specific sub-actions that are entailed by the surface realization. In order to formulate this condition, we make use of the support function $\mathcal{S}$, defined previously in definition (5.1), in section 5.6. Recall that $\mathcal{S}(M_i)$ expresses the communicative effectors that support the use of the mode M_i. For every matrix $A_i \in \Gamma$(*plan-ref*(e_i)), the following criterion must be satisfied:

$$\forall\, m_1, m_2, \tau \;\ni\; A_i[m_1, \tau] \neq 0,\; A_i[m_2, \tau] \neq 0,\; \mathcal{S}(m_1) \cap \mathcal{S}(m_2) = \emptyset \tag{6.6}$$

That is, at any given timestep, the communicative effectors that support any two modes in use must be disjoint.

6.3.3 The Δ Criterion

The Δ **criterion** stipulates that the elements of all matrices $A_i \in \Gamma$(*plan-ref*(e_i)) must be in accord with the minimum and maximum duration times specified for each of the mode-specific sub-actions. These were defined in (6.4) above. The duration times of each of the mode-specific

sub-actions in a given surface realization can be derived from the elements of the matrix. Recall from definition (6.1) that if the elements of the surface realization matrix $A[m_j, \tau] = l \ \forall \tau \in [t, t']$, this means that mode-specific sub-action $a_{j,l}$ is performed for the timesteps $t, \ldots, t'$.

For every matrix $A_i \in \Gamma(\textit{plan-ref}(e_i))$, the following criterion must be satisfied:

$\forall$ mode-specific sub-actions $a_{j,l} \ni A[m_j, \tau] = l, \ \tau \in [t, t']$ and $\{a_{j,l}, \delta_{\text{j,l,min}}, \delta_{\text{j,l,max}}\} \in \Delta_{M_j}$

$\delta_{j,l,\text{min}} \leq t' - t + 1 \leq \delta_{j,l,\text{max}}$

6.3.4 Completeness Criterion

The **completeness condition** stipulates that every semantic primitive in the communicative plan must be signalled. For an interlocutor to signal semantic primitive p_i, at least one mode-specific action $a_j \in \mathcal{R}(p_i)$ must be articulated (see definition (6.3) above). The formal statement of the completeness condition is as follows:

$$\forall p \in X, \ \exists \ s, t \ni A[s, t] = a, \text{ where } a \in \mathcal{A}_{|p} \tag{6.7}$$

This condition does not preclude the possibility that more than one mode-specific action might be used to signal a semantic primitive.

6.3.5 Preservation of Ordering Criterion

The completeness condition does not stipulate that the mode-specific actions should be articulated in any particular order. The **preservation of ordering condition** stipulates that the order in which the mode-specific actions for the semantic primitives are performed (as given by their times of onset[7]) must not be inconsistent with the partial ordering of the semantic primitives.

The formal statement of this condition is as follows:

if $\mathcal{O}_j$ is the partial order defined for semantic primitives X_j of the intended referent e_j, and $\mathcal{O}'$ is the partial ordering of the semantic primitives implied by the onset times of the mode-specific sub-actions, then

$$\nexists (x, y) \in \mathcal{O}' \ni (y, x) \in \mathcal{O}_j \tag{6.8}$$

6.3.6 Implementation

In the previous section, each of the various criteria was formalized as a expression that stipulates the mathematical relationships among the matrix rows, columns, and elements. These mathematical expressions allow us to characterize the derivation of $\Gamma(\textit{plan-ref}(e_i))$ as a constraint satisfaction problem (CSP), which, in turn, allows us to make use of existing solution-finding techniques. This is one of the chief benefits of characterizing surface realizations as matrices.

SCREAMER, a non-deterministic variant of Lisp, provides a mechanism for deriving all possible solutions to a given CSP. Each mathematical expression was implemented as a Lisp expres-

[7] Another partial ordering is given by the timesteps at which the mode-specific actions finish.

sion.[8] Thus, the task of deriving the set $\Gamma(\textit{plan-ref}(e_i))$ was formulated as the task of finding all matrices A that satisfy the criteria that were just described.[9]

The program has the following parameters: the mode repertoire, a set of semantic primitives, a partial ordering on the set of semantic primitives, the mode-specific sub-actions that signal each primitive, the durations of each mode-specific sub-action (the triples $\Delta_{M_j}\ \forall M_j \in \mathcal{M}$), and the maximum number of time steps. The input to the program is the communication plan. The SCREAMER program was subsequently ported to Java and implemented as a class.[10] The running time of the derivation process is exponential in the size of the surface realization matrices and the program is invoked once, upon start-up of MSIM.

Four examples of automatically-derived candidate surface realization matrices, $A_1, \ldots, A_4$, are shown in figure 6.1. This example is not intended to be realistic, but rather to demonstrate the formalism. Three modes of articulation were defined for agent C: gaze, vocalization, and gesture, which correspond to the rows 0, 1, and 2 of the matrices. No mode interferences were defined. Agent C has knowledge of three semantic primitives: $\{p_1 = \text{small}, p_2 = \text{red}, p_3 = \text{cube}\}$. The sub-actions that can potentially realize each of these primitives are listed in figure 6.1. The communicative plan $\textit{plan-ref}(e_6) = \{X_6, \mathcal{O}_6\}$ was specified, where $X_6 = \{p_1 = \text{small}, p_2 = \text{red}, p_3 = \text{cube}\}$ and the ordering relation $\mathcal{O}_6$ stipulates that $p_1 < p_2 < p_3$. These four matrices $A_1, \ldots, A_4$ are but a small subset of the entire set of automatically-derived candidate surface realizations for these parameter values. The multimodal communicative actions are described further in their accompanying captions.

[8] These were λ-expressions, small routines that take a single matrix as a parameter and return the boolean value of true or false to reflect whether the matrix satisfies the implemented criterion.

[9] An additional criterion was applied. The **basic economy of expression criterion** stipulates that for every time step, at least one mode-specific sub-action is being performed — i.e., that there should be no columns in the surface realization matrix A that contain only zeros:

$$\forall j = 1, \ldots, t, \exists i \ni A[i, j] \neq 0 \tag{6.9}$$

It is possible that other criteria may be needed for other types of surface realizations. Their identification and development is left for future work.

[10] Unlike SCREAMER, Java does not provide a built-in mechanism for CSP solution-finding and thus is not as ideal (additional functions needed to be implemented). However, the port was necessary for the sake of integration with the other components of MSIM.

Figure 6.1 Examples of automatically-generated surface realization matrices (see text and sub-figure captions for further description). A matrix element with value k in row i and column j indicates that the sub-action $a_{i,k}$ is performed at the timestep j. For instance, the value 2 in row 0 corresponds to sub-action $a_{0,2}$. The 0-elements are left blank for readability.

$$A_1 = \begin{matrix}\text{0: Gaze}\\ \text{1: Vocal.}\\ \text{2: Gest.}\end{matrix}\begin{bmatrix} & & & & & 2 & 2 & & \\ 1 & 1 & 1 & 1 & 1 & & & & \\ & & & & & & & 3 & 3 \end{bmatrix}$$

(a) The multimodal surface realization A_1 consists of a vocalization for p_1 ("small"), followed by a gaze action directed at the intended referent (i.e., coverage of the semantic primitive p_2 ("red")), and a iconic gesture for p_3 ("cube").

$$A_2 = \begin{matrix}\text{0: Gaze}\\ \text{1: Vocal.}\\ \text{2: Gest.}\end{matrix}\begin{bmatrix} & & & & & & & 2 & 2 & & \\ 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & & \\ & & & & & & & & & 3 & 3 \end{bmatrix}$$

(b) The multimodal surface realization A_2 consists of a vocalization for p_1 ("small"), followed by a vocalization for p_2 ("red"). A gaze action is directed toward the entity (coverage of the semantic primitive p_2, which begins slightly before the vocalization is complete). An iconic gesture for p_3 ("cube") is performed once the gaze is completed.

$$A_3 = \begin{matrix}\text{0: Gaze}\\ \text{1: Vocal.}\\ \text{2: Gest.}\end{matrix}\begin{bmatrix} & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & & \\ 1 & 1 & 1 & 1 & 1 & & & & & & 3 & 3 \end{bmatrix}$$

(c) The multimodal surface realization A_3 consists of a iconic gesture for p_1 ("small"), as well as a vocalization for p_1 (which begins slightly after the gesture begins). The vocalization of p_1 is followed by a vocalization for p_2 ("red"). An iconic gesture for p_3 ("cube") is performed once the vocalizations have been completed.

$$A_4 = \begin{matrix}\text{0: Gaze}\\ \text{1: Vocal.}\\ \text{2: Gest.}\end{matrix}\begin{bmatrix} & & & & & & & 2 & 2 & & & \\ & & & & & & & & & & & \\ 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \end{bmatrix}$$

(d) The multimodal surface realization A_4 consists of a two contiguous pointing gestures toward the intended referent (coverage of the semantic primitives p_1 ("small") and p_2 ("red")). The pointing gestures are also accompanied by gaze, also directed at the intended referent (additional coverage of the semantic primitive p_2). The gaze and gestures are followed by an iconic gesture for p_3 ("cube").

Sub-action	Mode Used	Primitive Signalled	Description
$a_{0,0}$	Gaze	$\emptyset$	directed at communication partner
$a_{0,1}$	Gaze	p_1	directed gaze
$a_{0,2}$	Gaze	p_2	directed gaze
$a_{0,3}$	Gaze	p_3	directed gaze
$a_{1,0}$	Vocal.	$\emptyset$	no action
$a_{1,1}$	Vocal.	p_1	dysarthric version of "small"
$a_{1,2}$	Vocal.	p_2	dysarthric version of "red"
$a_{1,3}$	Vocal.	p_3	dysarthric version of "cube"
$a_{2,0}$	Gest.	$\emptyset$	hands at rest
$a_{2,1}$	Gest.	p_1	iconic gesture for "small"
$a_{2,2}$	Gest.	p_1	pointing gesture
$a_{2,3}$	Gest.	p_1	iconic gesture for "cube"

The sub-actions that signal each semantic primitive are defined by:
$\mathcal{R}(p_1) = \{a_{0,1}, a_{1,1}, a_{2,1}\}$
$\mathcal{R}(p_2) = \{a_{0,2}, a_{1,2}, a_{2,2}\}$
$\mathcal{R}(p_3) = \{a_{0,3}, a_{1,3}, a_{2,3}\}$
The sub-actions that agent C is able to perform are defined by:
$\mathcal{A}(M_0) = \{a_{0,1}, a_{0,2}, a_{0,3}\}$
$\mathcal{A}(M_1) = \{a_{1,1}, a_{1,2}, a_{1,3}\}$
$\mathcal{A}(M_2) = \{a_{2,1}, a_{2,2}, a_{2,3}\}$

6.4 Analysis Technique for the Candidate Set

6.4.1 Overview

We now describe a technique for analyzing the set $\Gamma(\textit{plan-ref}(e_i))$. The goal of the analysis technique is to characterize a communicator's repertoire of mode strategies. In chapter 7, we will invoke MSIM under several different simulation conditions. By contrasting the simulation results, we will address the issue described in section 4.3.3: what is the effect of bottleneck reduction (which will be achieved through the use of a multimodal interface to a VOCA) on an interlocutor's repertoire of mode strategies?

In the first condition, agent C makes use of a unimodal VOCA; in the second, agent C makes use of a hypothesized multimodal VOCA (which entails the increase in mode conflict over a unimodal VOCA, but does not provide any additional benefit). In the third, agent C makes use of an hypothesized multimodal VOCA that implements bottleneck reduction (an increase in mode conflict, but with a concordant increase in the interpretability of the mode of synthesized speech, see section 4.3.2). (These middle conditions is needed in order to isolate the effect of increase mode conflict.)

We will characterize agent C's repertoire of mode strategies in a particular simulation condition on the basis of the candidate set $\Gamma(\textit{plan-ref}(e_i))$ that the agent architecture derives. This technique has two steps: the first component of the analysis is the identification of the mode strategy that is used in each of the candidate surface realizations. The second component, which is described in section 6.5, is the evaluation of each of the candidates. The evaluations of all of the candidates that make use of a particular mode strategy, collectively, serve to characterize that mode strategy.

6.4.2 Formalization of *Mode Strategy*

Recall, from chapter 4, the characterization of mode strategies as unimodal or multimodal (a *unimodal* mode strategy is the use of a particular mode in isolation, and a *multimodal* strategy is the use of two or more modes simultaneously, sequentially, or both simultaneously and sequentially). Given a surface realization matrix, determining which of these two strategies was used is straightforward. But this categorization of the types of mode strategies is too basic.

Several researchers have attempted to formalize the space of possible ways that modes might be used in more detail. For instance, Nigay and Coutaz [1993] identify two dimensions of mode use: the *temporal* and the *semantic*. For the temporal dimension, they distinguish between two modes being used "in sequence" and being used "in parallel". The two semantic interrelationships between mode-specific sub-actions are "in combination" and "independent". In the "combined" use of modes, the modes must be interpreted together, as each mode provides incomplete information [Bolt, 1984, cited by Smith et al. [1996]] (e.g., each provides context for the other and the interpretation of the actions in isolation will be inadequate). Each of these two dimensions has two possible values [Smith et al., 1996]; hence there are only four possible patterns of mode use. While this approach provides one basis for an inventory of mode strategies, it is too coarsely-grained and imprecisely defined. Instead, we will formulate mode strategy in an alternative way.

We believe that a better approach to the definition of different strategies of mode use is to define

Table 6.2 Possible values of $\nu(A_j)$, showing the different ways in which the modes might be used from the repertoire $\mathcal{M} = \{\text{GESTURE}, \text{VOCALIZATION}, \text{VOCA}\}$. The third column shows the mode strategy labels that were described in section 4.3.

$\nu(A_j)$	**Modes Employed in A_j**	**Type of Mode Strategy**[1]
001	Unaided mode of gesture	*unaided* (unaided-unimodal)
010	Unaided mode of vocalization	*unaided* (unaided-unimodal)
011	Multimodal strategy (modes of gesture and vocalization used in combination)	*unaided* (unaided-multimodal)
100	Aided mode of synthesized speech	*aided-unimodal*
101	Multimodal strategy (mode of synthesized speech accompanied by gesture)	*joint aided-unaided*
110	Multimodal strategy (mode of synthesized speech accompanied by vocalization)	*joint aided-unaided*
111	Multimodal strategy (modes of gesture, vocalization, and synthesized speech in combination)	*joint aided-unaided*

[1] The types as defined in section 4.3.

an equivalence relation on the set of candidate surface realization matrices, and then to define the repertoire of mode strategies on the basis of the different equivalence classes.

To start, we define $\sim$ as follows:

$$A_i \sim A_j \text{ iff } A_i \text{ and } A_j \text{ both realize the same semantic primitives in } \textit{plan-ref}(e_i). \tag{6.10}$$

We now define the relation $\dot{\sim}$ to characterize when surface realizations are similar with respect to their mode-specific sub-actions. We first define the function $\nu(A_j)$ that characterizes which modes are used in A_j; it is defined in (6.11) below:

$$\nu(A_j) = \texttt{Concatenate}(\, b_k, \ldots, b_1), \text{ where:} \tag{6.11}$$

$$b_i = \begin{cases} 1 & \text{if mode } M_i \text{ was used in } A_j \\ 0 & \text{if mode } M_i \text{ was not used in } A_j \end{cases}$$

For example, the possible values for the modes in the repertoire $\mathcal{M} = \{\text{GESTURE}, \text{VOCALIZATION}, \text{VOCA}\}$ are given in table 6.2. (In the simulations that will be described in chapter 7, the aided communicator has the same repertoire of three modes as in table 6.2.) We then define the relation $\dot{\sim}$:

$$A_i \dot{\sim} A_j \text{ iff } A_i \sim A_j \text{ (as defined in (6.10)) } \textit{and} \quad \nu(A_i) = \nu(A_j) \tag{6.12}$$

We define the function $\sigma(A_j)$ to be an indicator of the *degree of redundancy* in A_j; it returns the number of semantic primitives from $\textit{plan-ref}(e_i)$ that are multiply-signalled (see section 6.2.3 for a description). (The value of $\sigma(A_j)$ will range from $0, \ldots, |X_i|$.) Any referent in the simulations that will be described in chapter 7 can be signalled by three semantic primitives. Thus, $\sigma(A_j) \in$

Table 6.3 Labels for the equivalence sets corresponding to the equivalence relations $\dot{\sim}$ and $\ddot{\sim}$.

$\dot{\sim}$	$\ddot{\sim}$	
	$\sigma'(A_j) = 0$	$\sigma'(A_j) = 1+$
001	001-0	001-1+
010	010-0	010-1+
011	011-0	011-1+
100	100-0	100-1+
101	101-0	101-1+
110	110-0	110-1+
111	111-0	111-1+

$\{0, 1, 2, 3\}$. We define the function $\sigma'(A_j)$ to be a boolean indicator of whether any of the semantic primitives from *plan-ref*(e_i) are multiply-signalled. (The value of $\sigma'(A_j)$ is either 0, if no semantic primitives are multiply-signalled, or $1+$ if one or more semantic primitives are multiply-signalled).

We now define the relation $\ddot{\sim}$ to characterize when surface realizations are similar with respect to their mode strategy and degree of redundancy:

$$A_i \ \ddot{\sim} \ A_j \ \text{iff} \ A_i \sim A_j \ \text{(as defined in (6.10))} \ \textit{and} \\ \nu(A_i) = \nu(A_j) \ \textit{and} \ \sigma'(A_i) = \sigma'(A_j) \tag{6.13}$$

Any equivalence relation defined on the set of surface realizations gives rise to equivalence classes that fully partition the set of candidate surface realizations for a given *plan-ref*(e_i); we define agent C's repertoire of mode strategies on the basis of the subsets yielded by the equivalence relation $\ddot{\sim}$. We can use the values of $\sigma'(A_j)$ and $\nu(A_j)$ *together* to label each of the subsets (or mode strategies). Table 6.3 gives the labels for the 14 different equivalence classes that are derived by the equivalence relation $\ddot{\sim}$.[11]

6.5 Candidate Evaluation

6.5.1 Motivation

In the context of its decision-making process, agent C evaluates each of the candidate multimodal surface realizations with respect to two different attributes: understandability and physical effort. The agent architecture then uses a value function in order to synthesize these two evaluations into an overall evaluation. Agent C then chooses the "best" candidate. Agent C's choice is an output of the simulation. But also of interest are the evaluations of all the other candidate surface realizations. As described above, these candidates can be partitioned according to mode strategy, and the collective evaluation of all the surface realizations that make use of a particular mode

[11] In a previous series of MSIM simulations, the equivalence relation $\ddot{\sim}$ was used:

$$A_i \ \ddot{\sim} \ A_j \ \text{iff} \ A_i \sim A_j \ \text{(as defined in (6.10))} \ \textit{and} \\ \nu(A_i) = \nu(A_j) \ \textit{and} \ \sigma(A_i) = \sigma(A_j) \tag{6.14}$$

This equivalence relation yields 28 different equivalence classes, which proved to be unwieldy. The equivalence relation $\ddot{\sim}$ replaced $\ddot{\sim}$.

strategy is a reflection on the mode strategy itself.

We can determine the effect of a particular VOCA (unimodal or multimodal) on agent C's repertoire of mode strategies by linking the mode strategy categorizations with the candidate evaluations. Thus, the evaluations of all of the candidate surface realizations, broken down by mode strategy used, is also considered an output of the simulation.

6.5.2 State Transition

As stated in section 5.5, the agent architecture in MSIM implements *decision under certainty*, meaning there is exactly one possible consequence for each alternative — the choice among alternatives is equivalent to a choice among consequences. MSIM makes use of a state transition model, which determines the consequence state, given a current state and a particular action. The model is represented by a function of the form given by g below.

$$g(S_i, A_j) = S_j \tag{6.15}$$

Without the assumption that there is a one-to-one mapping between the performance of an action and the state that follows from its performance, MSIM would require probability distributions over sets of possible actions; these distibutions would need to be based on empirical data that does not presently exist and its collection is outside the scope of this dissertation. Instead, this simplying assumption allows us to focus on developing techniques for demonstrating the effect of the parameter values on agent C's set of candidates.

Recall that the state space S for the multimodal referential communication task was defined in section 5.5. The simplying assumption means that we define $\forall A_i \in \Gamma,\ \exists\ S_i \in S\ \ni\ p(S_i|A_i) = 1$. When choosing from among the candidate surface realizations $\Gamma = \{A_1, \ldots, A_n\}$, the agent evaluates each in terms of the value of the state that follows from it. As also described in section 5.5, the decision maker's preferences are simulated by a single-attribute or multi-attribute *value function*; the function introduces an ordering on the set of consequences and thus also ranks the alternatives. The value function will be described in section 6.5.5. In MSIM, agent C chooses the alternative that results in the consequence with the highest value. And, as described in the previous sub-section, the values of all the candidate surface realizations, broken down by mode strategy used, is also considered an output of the simulation.

We describe below the state transition model, which builds upon the definitions of a state in the MMRC task and the space of possible states that was described earlier in section 5.5.

6.5.3 State Transition with Respect to the Procedural Goal Attainment

In step 2 of the MMRC task, agent C derives a surface realization for its communication plan in consideration of its procedural goal, which is to perform the task with as little effort as possible. Recall from section 5.5 that the state of the MMRC task when agent C begins step 2 is $S_1 = \langle 0, 1 \rangle$.[12] The agent architecture calls upon the state transition model in step 2 to determine, given a candidate

[12]The first component is 0, which means that the domain goal of being understood is not yet satisfied (since no communicative action has yet been produced), and the second component is 1, which means that the procedural goal of exerting as little physical effort as possible remains fully satisfied.

surface realization A_i, the amount of physical effect that the performance of A_i will entail. The architecture does this for each of the candidate surface realizations.

We assume a model in which the physical cost of a communicative action is derived from the physical cost of signalling each of the semantic primitives. Since each semantic primitive p_i can be signalled by one or more mode-specific sub-actions, we hypothesize that the cost of signalling p_i depends on which mode-specific sub-action (or actions) is (are) used to signal it. We further hypothesize that the cost of a mode-specific sub-action can be modelled in terms of the *cost* of the mode used for it and an additional *scaling factor* that is specific to the particular mode- and semantic-primitive configuration.

We define $\mathfrak{M}_{k,p_i} = \{\mu_1, \ldots, \mu_s\}$ to be the set of all mode-specific sub-actions μ in A_k that serve to signal the semantic primitive p_i. Given a mode-specific sub-action μ, we denote the semantic primitive that it signals by $p_{i|\mu}$ and the mode that is used by $M_{j|\mu}$. This notation will be used in subsequent definitions. We will now develop an approach to deriving the cost of a mode-specific sub-action μ.

We hypothesize that each mode $M_j \in \mathcal{M}$ is characterized by a *cost index* and, to represent these indices, we define the *mode repertoire cost* function C_1 in (6.16) below.

$$C_1 : \mathcal{M} \longrightarrow (0, 1] \tag{6.16}$$

A cost index such as $C_1(M_j) = 0$ (if it were permitted) would signify that there is no cost associated with the use of the mode M_j. Since, in practice, every mode requires at least some effort, the value is disallowed. One of the effects of physical disorder on an individual who has a communication disorder is the amplification of the costs of the modes in the individual's repertoire. The higher the index of a mode, the more physically costly it is to use it. An index of 1 signifies that the mode is maximally expensive, which is defined here to mean that if the interlocutor articulates a mode-specific sub-action with it, then he or she would experience a level of fatigue that would be detrimental to the production of subsequent communicative actions.

We further hypothesize that the cost of a mode-specific sub-action also depends on the fit between the semantic primitive that is signalled by it and the mode that is used to do so. We model this fit by a set of scaling factors for the mode costs. The scaling factors are intended to capture the intuition that, for each semantic primitive, certain modes are more costly for signalling it than others. Whereas the mode costs characterize the physical communicative abilities of the interlocutor, the scaling factors characterize the ability of the interlocutor *with respect to* the domain $\mathcal{X}$ of semantic primitives. We define the *semantic primitive cost scaling* function C_2, given in (6.17) below, to represent the scaling factors.

$$C_2 : \mathcal{X} \times \mathcal{M} \longrightarrow [1, \infty) \tag{6.17}$$

We define the scaling factors to be values that can serve *only* to *increase* the cost of performing a mode specific sub-action — a mode's cost index cannot be reduced by the type of semantic primitive that is signalled by it. As the value of the scaling factor increases from 1, the greater the physical effort that is required to signal semantic primitive p_i using the mode M_j.

Given the functions $C_1(M_j)$ and $C_2(p_i, M_j)$, we now define the *cost* of a mode-specific sub-action μ to be the product of the cost of the mode $M_{j|\mu}$ and the scaling factor that is associated with

$M_{j|\mu}$ and $p_{i|\mu}$. We denote this function by $\alpha_{\mathcal{A}}(\mu)$ and its formula is given in (6.18) below.[13]

$$\alpha_{\mathcal{A}}(\mu) = C_2(p_{i|\mu}, M_{j|\mu}) \cdot C_1(M_{j|\mu}). \tag{6.18}$$

We now use $\alpha_{\mathcal{A}}$ as the basis for two approaches to derive the cost of a group of mode-specific sub-actions $\mathfrak{M}_{k,p_i}$. The first approach makes use of an additive model (the subscript of α is $\mathfrak{M}$ rather than $\mathcal{A}$, to denote that the function's domain is a set of mode-specific sub-actions rather than a single one):

$${}_1\alpha_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}) = \sum_{\mu \in \mathfrak{M}_{k,p_i}} C_2(p_{i|\mu}, M_{j|\mu}) \cdot C_1(M_{j|\mu}) \tag{6.19}$$

In the second approach, the cost of a group of mode-specific sub-actions also reflects the differences in cost of signalling different numbers of mode-specific sub-actions. If only one mode-specific action is used to signal a given semantic primitive p_i, then no cost penalty is applied. If the primitive is multiply-signalled, then penalties are applied.

$${}_2\alpha_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}) = \pi_{\alpha}(\mathfrak{M}_{k,p_i}) \sum_{\mu \in \mathfrak{M}_{k,p_i}} C_2(p_{i|\mu}, M_{j|\mu}) \cdot C_1(M_{j|\mu}) \tag{6.20}$$

The function $\pi_{\alpha}(\mathfrak{M}_{k,p_i})$ is the penalty function, and is defined in (6.21) below.[14]

$$\pi_{\alpha}(\mathfrak{M}_{k,p_i}) = \begin{cases} 1 & \text{if } |\mathfrak{M}_{k,p_i}| = 1 \\ 1 + \gamma(\mathfrak{M}_{k,p_i})(\pi' - 1) & \text{if } |\mathfrak{M}_{k,p_i}| = 2 \\ \pi' + \gamma(\mathfrak{M}_{k,p_i})(\pi'' - \pi') & \text{if } |\mathfrak{M}_{k,p_i}| = 3 \end{cases} \tag{6.21}$$

In the penalty function π_{α}, we make use of the *simultaneity function* $\gamma(\mathfrak{M}_{k,p_i})$. The function determines the proportion of the mode-specific sub-actions in $\mathfrak{M}_{k,p_i}$ that are performed simultaneously (it looks at all of the pairwise combinations of the mode-specific sub-actions in $\mathfrak{M}_{k,p_i}$ and counts those that are performed with non-null temporal overlap). If two mode-specific sub-actions are used to signal a semantic primitive, but they are not performed simultaneously, then no penalty is applied. The penalty is scaled up according to the degree they are performed simultaneously, with the maximum penalty being π' (i.e., when $\gamma(\mathfrak{M}_{k,p_i}) = 1$). If three mode-specific sub-actions are used to signal a semantic primitive, but they are not performed simultaneously, then the penalty value is $\pi_{\alpha} = \pi'$ (i.e., the cost of performing two sub-actions simultaneously has been equated with the cost of performing three sub-actions sequentially). The penalty is scaled up according to the degree they are performed simultaneously, with the maximum penalty being π'' (i.e., when $\gamma(\mathfrak{M}_{k,p_i}) = 1$). Thus, the values π' and π'' characterize the additional burden of producing coordinating and multiple mode-specific sub-actions.

In MSIM, the constants π' and π'' are parameters that are passed to the cost model. In this work, a number of values were used; however it was found that as long as $1 < \pi' < \pi''$, the same

[13]The subscript $\mathcal{A}$ for this function is derived from the notation for each semantic primitive p_i's inventory of mode-specific sub-actions, which is denoted by $\mathcal{A}(M_i)$ (see section 6.2.4).

[14]In the definition (6.21), the entity $|\mathfrak{M}_{k,p_i}|$ gives the number of mode-specific sub-actions that signal semantic primitive p_i. If $|\mathfrak{M}_{k,p_i}| = 1$, the semantic primitive p_i is signalled by one mode-specific sub-action, and if $|\mathfrak{M}_{k,p_i}| > 1$, the semantic primitive p_i is signalled by multiple mode-specific sub-actions (i.e., it is *multiply-signalled*).

relative effect on the different mode strategies occurred. The same basic patterns were seen, and the differences between different sets of π' and π'' values concerned the magnitude of the effects on the mode strategy evaluations. For the simulation conditions that will be described in the next chapter, the values of $\pi' = 2$ and $\pi'' = 3$ were used (to correspond with the intuition that it is twice as costly to perform two mode-specific sub-actions simultaneously and three times as costly to perform three).

Finally, we now use ${}_1\alpha_{\mathfrak{M}_{k,p_i}}$ and ${}_2\alpha_{\mathfrak{M}_{k,p_i}}$ as the bases for deriving the cost of a multimodal surface realization, A_k in two different ways. The general form is given by:

$$\alpha(A_k) = \sum_{p_i \in X_r} \alpha_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}). \tag{6.22}$$

The summation in (6.22) ranges over all of the semantic primitives signalled in A_k and derives the sum of the costs of the groups of mode-specific sub-actions that signal the semantic primitives in *plan-ref*(e_r). The simulation tool MSIM is able to make use of either of the two functions ${}_1\alpha_{\mathfrak{M}}(\mathfrak{M}_{k,p_i})$ or ${}_2\alpha_{\mathfrak{M}}(\mathfrak{M}_{k,p_i})$; the two resulting cost functions are denoted by α_1 and α_2.

Last, we use the derived cost value in order to derive a consequence state attribute $s_P \in [0,1]$. We assume that the degree to which the procedural goal is satisfied is negatively correlated with the amount of physical effort required to perform the communicative action A_i. We distinguish two different state transition functions: $g_1 : S \times A \rightarrow S$, which is based on α_1, and $g_2 : S \times A \rightarrow S$, which is based on α_2, defined in (6.34) and (6.35) below. In these functions, we make use of the values $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ instead of α_1 and α_2, respectively.[15] For the simulations described in the next chapter, the values $\texttt{min}(\alpha_1)$, $\texttt{max}(\alpha_1)$, $\texttt{min}(\alpha_2)$, and $\texttt{max}(\alpha_2)$ are reported.

$$g_1(\langle 0,1\rangle, A_i) = \langle 0, 1 - \tilde{\alpha}_1(A_i)\rangle \tag{6.23}$$

$$g_2(\langle 0,1\rangle, A_i) = \langle 0, 1 - \tilde{\alpha}_2(A_i)\rangle \tag{6.24}$$

6.5.4 State Transition with Respect to the Domain Goal Attainment

In the MMRC task, the interlocutor's domain goal is to establish the identity of the intended referent e_r to agent L. Thus, given a candidate surface realization A_i, one of the tasks of the state transition model is to determine the degree to which this goal has been met. The state transition model is called upon to perform this task when agent C is performing step 2 of the MMRC task (although the state transition does not actually take place until step 3 has been completed). Agent C evaluates the attribute s_D of the consequence states for all possible actions when in performing step 2 of the MMRC task.[16]

The ideal consequence state that one could imagine following from the state $S_2 = \langle 0, x\rangle$ is the state $S_3 = \langle 1, x\rangle$; the value s_D is updated from 0 to 1, meaning the domain goal has been fully

[15]The value of $\tilde{\alpha}_1$ is based on the value of α_1 (which falls within the interval $[\texttt{min}(\alpha_1), \texttt{max}(\alpha_1)]$), scaled so that it falls within the interval $[0,1]$. The value of $\tilde{\alpha}_2$ was derived analogously. This scaling was required so that the differences $1 - \tilde{\alpha}_1(A_i)$ and $1 - \tilde{\alpha}_2(A_i)$ would always be positive.

[16]Observe that the value s_P in $\langle s_D, s_P\rangle$ does not change during step 3. In this step, agent L attends to agent C's communicative action and interprets it. In MSIM, when agent C considers the consequences of the alternative surface realizations $\{A_1, \ldots, A_n\}$, it considers the possible consequence states that follow step 3 directly.

achieved (i.e., agent L's interpretation of the intended referent is exactly the referent that agent C intended; the value $s_P = x$ remains unchanged from S_2 to S_3). Although this outcome is possible, it is but one of many. Other outcomes correspond to consequences in which the communication partner misidentifies the intended referent. We hypothesize that it is possible to derive a degree of misidentification (e.g., if the surface realization "small blue cube" gets misinterpreted as referring to the entity that is a large red ball or the entity that is a small blue ball, the degree of the first misidentification is greater than the second).

We hypothesize that the degree to which the identity of e_r is correctly interpreted by L depends on the degree to which each of the semantic primitives $X_r = \{p_1, \ldots, p_{n_r}\}$ that are associated with it is correctly interpreted. An intended referent can be identified correctly only if all of the semantic primitives that distinguish it from the other competitor candidates are successfully interpreted. If some, but not all, of the primitives are not clearly signalled, the communication partner will be left with an incomplete interpretation. (However the partner might still correctly identify e_i by a lucky guess from among the set of possible matches.)

As in the case of the cost model of multimodal communicative action, we assume a model in which the *interpretability* of the action with surface realization A_k is derived from the interpretability of each of the semantic primitives that the action's surface realization serves to signal. We hypothesize that the interpretability of a mode-specific sub-action that signals semantic primitive p_i depends on the mode (or modes) that is (are) used to signal it. We will define the function R_1 and R_2 below, and define the *interpretability* of a mode-specific sub-action μ by:

$$\beta_{\mathcal{A}}(\mu) = 1 - R_2(p_{i|\mu}, M_{j|\mu}) \cdot R_1(M_{j|\mu}). \tag{6.25}$$

We hypothesize that each of mode $M_j \in \mathcal{M}$ can be characterized by an *interpretability difficulty index*; this index is represented by the value of $R_1(M_j)$ in (6.25) above. This index reflects the amount of shared background with C that L requires in order to interpret C's actions with respect to that mode. We hypothesize that this is another mechanism whereby physical disorders have their impact on the production of communicative actions — they can increase the amount of background that partners must share with an individual in order to interpret successfully his or her communicative actions. An index value of 0 for a mode M_k signifies that that mode is maximally interpretable (i.e., any communication partner from the aided communicator's community would have enough background knowledge to interpret actions signalled with respect to the mode). For instance, almost all communication partners are able to interpret the mode of synthesized speech.[17] In this case, the amount of common background that is required between C and L is low and thus the index of difficulty for the aided mode would be low (say 0.1; one in ten would misinterpret the synthesized speech). But other modes, such as gesture or vocalization, typically cannot be interpreted by unfamiliar communication partners. Vocalizations might be unintelligible to everyone except those who are very familiar to the aided communicator. For the modes of gesture and vocalization, the amount of required common background between the aided communicator and the communication partner would be high (and thus the index would be high). An

[17]The intelligibility of synthesized speech depends on the experience level of the communication partner; the best DECTalk and MacinTalk voices ("Paul" and "Bruce", respectively) have at best sub-90% intelligibility in the best cases [Hustad et al., 1998].

index value of 1 for a mode M_k signifies that that mode is maximally uninterpretable (i.e., the actions produced using this mode are not interpretable by anyone). The *mode repertoire interpretability difficulty* function $R_1(M_j)$ is defined in (6.26) below.

$$R_1 : \mathcal{M} \longrightarrow [0, 1] \tag{6.26}$$

We further hypothesize that the interpretability of a mode-specific sub-action also depends on the appropriateness of the mode with which a semantic primitive is being signalled. For example, a semantic primitive might be misunderstood if imprecise hand gestures are used to signal it rather than if synthesized speech were to be used. The characteristics of the mode that are specific to a given semantic primitive are represented by the values of the *semantic primitive interpretability scaling* function $R_2(p_i, M_j)$, defined in (6.27) below.

$$R_2 : \mathcal{X} \times \mathcal{M} \longrightarrow [0, k), \text{ where the value of } k \text{ depends on } R_1 \tag{6.27}$$

The function values represent the scaling factors for each of the semantic primitives with respect to each of the modes in the interlocutor's repertoire. The primitive-specific scaling factor of 1.0 leaves $R_1(M_j)$ unaffected and is used to represent the baseline case. A scaling factor on the interval $(0, 1)$ further reduces a mode's interpretability difficulty index (and indicates that the semantic primitive is especially amenable to being signalled with respect to that mode). A scaling factor of 0 stipulates that the interpretability of a mode-specific sub-action for semantic primitive p_i using mode M_j is 0 (and indicates that the semantic primitive that is thus signalled can be interpreted by any communication partner). The value of the maximum possible scaling factor is k, which depends on R_1, as we will now explain. Scaling factors on the interval $(1, k), k > 1$ increase a mode's interpretability difficulty index; such scaling factors are used to indicate that semantic primitives with respect to that mode are more difficult to interpret successfully than if they were to be signalled using other modes (namely the baseline case, described above). In the extreme, the scaling factor k chosen such that $k \cdot R_1(M_j) = 1$ indicates that p_i, when signalled using mode M_j, would not be interpretable by anyone. However, the value k must be chosen with care, since $k \cdot R_1(M) \leq 1$, $\forall M \in \mathcal{M}$ in order to ensure that the value of $\beta_{\mathcal{A}}(\mu)$ is non-negative.

The product $R_1(M_j) \cdot R_2(p_i, M_j)$ gives a interpretability *difficulty* value for the mode-specific sub-action μ. The value of the term $1 - R_1(M_j) \cdot R_2(p_i, M_j)$ provides a measure of the degree to which the mode-specific sub-action μ can be interpreted successfully.

Similarly to the previous section, we define two ways of deriving the *interpretability* of a group of mode-specific sub-actions $\mathfrak{M}_{k,p_i}$. The functions are given in (6.29) and (6.30) below. Then we define the interpretability of a surface realization A_k in terms of the interpretability of each of the semantic primitives. The general form is given in (6.28) below.

$$\beta(A_k) = \sum_{p_i \in X_i} \beta_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}) \tag{6.28}$$

The summation in (6.28) ranges over all of the semantic primitives signalled in A_k and derives the sum of the intepretability values of the groups of mode-specific sub-actions that signal the

semantic primitives in *plan-ref*(e_r). The simulation tool MSIM is able to make use of either of the two functions ${}_1\beta_{\mathfrak{M}}(\mathfrak{M}_{k,p_i})$ and ${}_2\beta_{\mathfrak{M}}(\mathfrak{M}_{k,p_i})$; the two resulting interpretability functions are denoted by β_1 and β_2.

The first function ${}_1\beta_{\mathfrak{M}}$ is an additive model of the interpretabilities of the mode-specific sub-actions:

$$ {}_1\beta_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}) = \sum_{\mu \in \mathfrak{M}_{k,p_i}} 1 - R_2(p_{i|\mu}, M_{j|\mu}) \cdot R_1(M_{j|\mu}) \tag{6.29} $$

The second function ${}_2\beta_{\mathfrak{M}}$ takes into account the synergistic effect of simultaneously-performed mode-specific sub-actions:

$$ {}_2\beta_{\mathfrak{M}}(\mathfrak{M}_{k,p_i}) = \rho_\beta(\mathfrak{M}_{k,p_i}) \sum_{\mu \in \mathfrak{M}_{k,p_i}} 1 - R_2(p_{i|\mu}, M_{j|\mu}) \cdot R_1(M_{j|\mu}) \tag{6.30} $$

The *reward* function $\rho_\beta(\mathfrak{M}_{k,p_i})$ is defined in (6.31) below:[18]

$$ \rho_\beta(\mathfrak{M}_{k,p_i}) = \begin{cases} 1 & \text{if } |\mathfrak{M}_{k,p_i}| = 1 \\ 1 + \gamma(\mathfrak{M}_{k,p_i})(\pi' - 1) & \text{if } |\mathfrak{M}_{k,p_i}| = 2 \\ \pi' + \gamma(\mathfrak{M}_{k,p_i})(\pi'' - \pi') & \text{if } |\mathfrak{M}_{k,p_i}| = 3 \end{cases} \tag{6.31} $$

The function was designed to derive higher rewards for situations in which a semantic primitive is signalled by multiple mode-specific sub-actions than when it is signalled by a single mode-specific sub-action. The intuition is that if a semantic primitive is multiply-signalled, the communication partner might be able to exploit this in order to derive a interpretation that is better than what would be possible if it were signalled by single mode-specific sub-action. We designed the function so that it first rewards for the number of mode-specific sub-actions used, and then for simultaneous performance of those actions. The reward function ρ_β also makes use of the *simultaneity function* $\gamma(\mathfrak{M}_{k,p_i})$ (described previously for the penalty function π_α). If two mode-specific sub-actions are used to signal a semantic primitive, then the degree to which the actions are performed simultaneously determines the reward (the maximum reward is π' if they are completely simultaneous). If three mode-specific sub-actions are used, the reward ranges from π' to π'', scaled according to their degree of simultaneity.

The reward function ρ_β and the the penalty function π_α were specifically designed so that they both make use of the values of π' and π'' — this way, the magnitude of the maximum reward is symmetric with the maximum penalty.

Last, we derive the consequence state attribute $s_D \in [0, 1]$ on the basis of the interpretability value of action A_k. We define additional state transitions for the functions g_1 and g_2 that were defined in the previous section. Analogously to the case in the previous section, in these functions we make use of the values $\tilde{\beta}_1$ and $\tilde{\beta}_2$ instead of β_1 and β_2, respectively.[19] For the simulations

[18] Recall from earlier definitions that the entity $|\mathfrak{M}_{k,p_i}|$ in definition (6.31) gives the number of mode-specific sub-actions that signal semantic primitive p_i.

[19] The value of $\tilde{\beta}_1$ is based on the value of β_1 (which falls within the interval $[\min(\beta_1), \max(\beta_1)]$), scaled so that it falls within the interval $[0, 1]$. The value of $\tilde{\beta}_2$ was derived analogously. This scaling was required so that the value $\tilde{\beta}_1(A_i)$ and $\tilde{\beta}_2(A_i)$ would always be equal to or less than 1 (other values would result in states that are not defined in our state space).

described in the next chapter, the values $\texttt{min}(\beta_1)$, $\texttt{max}(\beta_1)$, $\texttt{min}(\beta_2)$, and $\texttt{max}(\beta_2)$ are reported.

$$g_1(\langle 0, s_P \rangle, A_i) = \langle \tilde{\beta}_1(A_i), s_P \rangle \tag{6.32}$$

$$g_2(\langle 0, s_P \rangle, A_i) = \langle \tilde{\beta}_2(A_i), s_P \rangle \tag{6.33}$$

Thus, we have two different functions that define the transitions from the state at the start of step 2 to the state once step 3 has been completed (collapsing two state transitions, from S_1 to S_2 and from S_2 to S_3, into one transition — see section 5.5 for a description of the states):

$$g_1(\langle 0, 1 \rangle, A_i) = \langle \tilde{\beta}_1(A_i), 1 - \tilde{\alpha}_1(A_i) \rangle \tag{6.34}$$

$$g_2(\langle 0, 1 \rangle, A_i) = \langle \tilde{\beta}_2(A_i), 1 - \tilde{\alpha}_2(A_i) \rangle \tag{6.35}$$

6.5.5 The Value Function

When designing their multimodal communicative actions, human communicators find a way to balance the competing goals of being understood as intended and minimizing the expenditure of physical effort. In MSIM, this is implemented by agent C's use of a *value function*, which evaluates the value of the state that follows the performance of a multimodal communicative action. The value function combines the domain-goal-specific and procedural-goal-specific values of a consequence state S_i, which will be represented by $V_D(S_i)$ and $V_P(S_i)$, respectively. From the definition of the consequence states above, we can see that these function values appear directly in the representation of the state (i.e., $V_D(S_i) = s_D$, $V_P(S_i) = s_P$). The *overall* value of a state is a weighted sum of the two attribute-specific values:

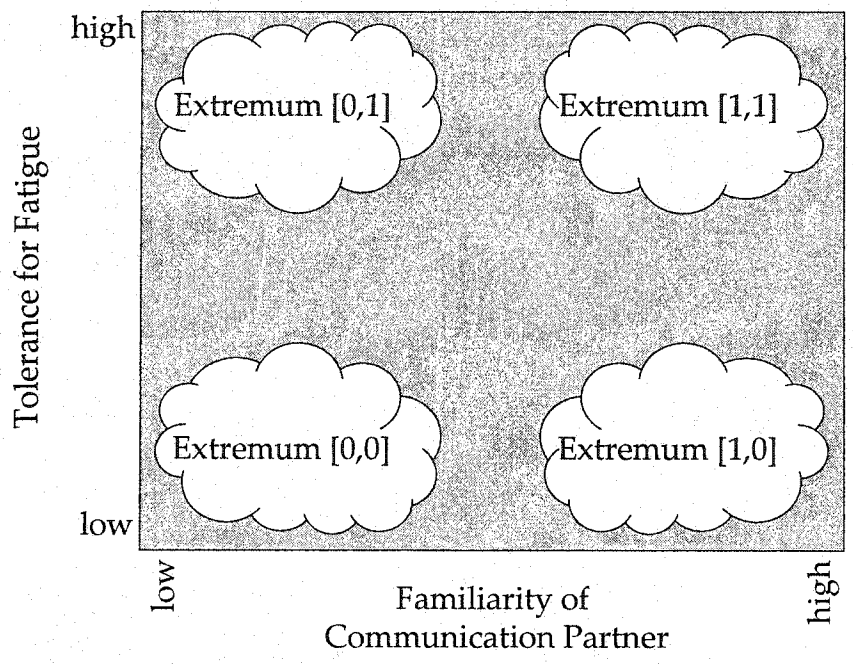
$$V_j(S_i) = w_D \cdot V_D(S_i) + w_P \cdot V_P(S_i). \tag{6.36}$$

Notice that the value function V_j is indexed. MSIM makes use of the assumption that relative weights of the attribute-specific values varies from one communicative context to the next (possibly even from one invocation of the production process to the next), albeit in a principled way. The index j identifies the particular value function that is to be used in a given simulation condition.

In MSIM, the weights vary as a function of two factors: agent C's *tolerance for fatigue* and the *familiarity of the communication partner*. These factors are parameters of a simulation condition, and each is assigned a value from $[low, med, high]$ (see section 5.6). We illustrate the four extrema of the pairwise combination of these two dimensions in figure 6.2 and describe below the hypothesized weights in the value function that might be used in each of these scenarios. We will adopt the convention that the values are normalized and that $w_D + w_P = 1$.

Extremum $[high, high]$ represents the "ideal" situation for the aided communicator: the partner is familiar and the communicator has a high tolerance to fatigue. Familiar communication partners are able to interpret successfully all sorts of multimodal communicative actions (and not just those actions which rely heavily on the mode of synthesized speech). In this case, we hypothesize that the communicator will weigh most heavily the procedural goal, perhaps even to the exclusion of the domain goal (since all of the candidate multimodal surface realizations meet the criteria for

Figure 6.2 An overview of the extrema of the pairwise combination of these two dimensions.



satisfactorily implementing the communication plan).

Thus, the relative importance of the goals would be given by $w_D = 0$ and $w_P = 1$. We denote the value function that would be used in this situation and with these weights by $V_{\mathrm{H,H}}$. Extremum $[high, low]$ also represents a situation in which the partner is familiar but here the aided communicator's tolerance for fatigue is low. In this scenario, the aforementioned strategy would also be the best, thus, the relative importance of the goals would also be given by $w_D = 0$ and $w_P = 1$. We denote the value function that would be used in this situation by $V_{\mathrm{H,L}}$.

Extremum $[low, high]$ represents a situation in which the aided communicator has a partner who is unfamiliar, but the communicator's tolerance to fatigue is high. Extra effort can be expended on the performance of multimodal communicative actions that are designed to be highly interpretable (such as those actions which rely heavily on the aided mode of synthesized speech). We hypothesize that the aided communicator's strategy is to give the domain goal the highest priority and the procedural goal little or no priority: the weights of the goal-specific attributes would be given by $w_D = 1$ and $w_P = 0$. We denote the value function that would be used in this situation by $V_{\mathrm{L,H}}$.

Extremum $[low, low]$ represents the worst situation for an aided communicator: the partner is unfamiliar and the communicator has a low tolerance to fatigue. It is extremely difficult, if not impossible, for the aided communicator to satisfy both the domain and the procedural goals satisfactorily; surface realizations that are low-effort also are those that are most difficult for an unfamiliar partner to interpret successfully. We hypothesize that the aided communicator's strategy is to attempt to satisfy both goals as much as is possible and not to weight one more than the other. In this case, the weights of the goal-specific attributes would be given by $w_D = w_P = 0.5$. We denote

Table 6.4 A summary of the different value functions used in MSIM, given as a function of the parameter values for *communication partner familiarity* and *tolerance for fatigue*. The first row of each cell gives the name of the value function used (and the corresponding index in parenthesis), the second gives the weights for the goal-specific attributes.

tolerance to fatigue	*communication partner familiarity*		
	low	med	high
low	$V_{\text{L,L}}$ (V_3) $w_D = 0.50, w_P = 0.50$	$V_{\text{M,L}}$ (V_5) $w_D = 0.250, w_P = 0.750$	$V_{\text{H,L}}$ (V_6) $w_D = 0.00, w_P = 1.00$
med	$V_{\text{L,M}}$ (V_2) $w_D = 0.75, w_P = 0.25$	$V_{\text{M,M}}$ (V_4) $w_D = 0.375, w_P = 0.625$	$V_{\text{H,M}}$ (V_6) $w_D = 0.00, w_P = 1.00$
high	$V_{\text{L,H}}$ (V_1) $w_D = 1.00, w_P = 0.00$	$V_{\text{M,H}}$ (V_3) $w_D = 0.500, w_P = 0.500$	$V_{\text{H,H}}$ (V_6) $w_D = 0.00, w_P = 1.00$

the value function that would be used in this situation by $V_{\text{L,L}}$.

The agent architecture makes use of the parameter values of *communication partner familiarity* and *tolerance for fatigue* in order to determine which value function should be used. Table 6.4 lists all of the pairwise combinations of these two parameters and the values functions that correspond to those simulation conditions. The values for functions $V_{\text{L,L}}$, $V_{\text{H,L}}$, $V_{\text{L,H}}$, and $V_{\text{H,H}}$ were derived from extrema $[low, low]$, $[high, low]$, $[low, high]$, and $[high, high]$, respectively. The values for all the other functions were derived by interpolation.

If we were to rank the value functions according to the weight given to the domain goal relative to the procedural goal, then the order would be as given in (6.37). The indices for the value functions are derived from this order and are listed in (6.38) and also shown in parenthesis in table 6.4. These indices are used in chapter 7 as the values for the x-axis of scatterplots that illustrate the simulation results.

$$V_{\text{L,H}} < V_{\text{L,M}} < V_{\text{L,L}}, V_{\text{M,H}} < V_{\text{M,M}} < V_{\text{M,L}} < V_{\text{H,L}}, V_{\text{H,M}}, V_{\text{H,H}} \tag{6.37}$$

$$V_1 < V_2 < V_3 < V_4 < V_5 < V_6 \tag{6.38}$$

6.6 Summary

This chapter provided a detailed description of the process whereby a set of candidate surface realizations for a communication plan of the form *plan-ref*(e_i) is derived, evaluated, and the "best" surface realization is selected.

Section 6.2 described the architecture of the multimodal surface realization module and provided pseudo-code for this process of deriving a communication plan and a multimodal surface realization for it. This section concluded with a description of the representation of candidate surface realizations using matrices.

Section 6.3 described the mechanism for generating a set of candidate surface realizations for a given communication plan. Two different categories of different criteria were described. The first type formalized the condition that a candidate surface realization must be performable by the

agent, given its available communicative resources. The second type formalized the condition that a candidate must adequately realize the underlying communication plan. The implementation of this algorithm and some example outputs were provided.

Section 6.4 described a technique for categorizing the mode strategy that corresponds to each candidate surface realization. Mode strategies are distinguished on the basis of the modes used and the degree of redundancy that is entailed. This technique is needed so that we can determine computationally the repertoire of mode strategies that is available to agent C in a given simulation condition. This is done on the basis of the candidate set that is generated for that condition. Moreover, we want to characterize each of the mode strategies; this was described in the subsequent section, section 6.5. This evaluation is done by the agent architecture in the context of its decision-making process — in order for the agent to perform its task, it must identify the best candidate from among all of the candidates. But another use of the candidate evaluations was described. Once the candidates are categorized according to mode strategy used, the evaluations of the candidates that correspond to each mode strategy, collectively, serve to characterize the mode strategy. Section 6.5 described the state transition model, which hypothesizes the values of two attributes of the state that would follow as a consequence of the agent performing a given candidate surface realization. One of the attributes concerns the amount of physical effort expended, which is connected to the satisfaction of the agent's procedural goal, and the other concerns the interpretability of the communicative action, which is connected to the satisfaction of the agent's domain goal. In the last component of this section, a set of six different value functions was derived and discussed. Each of these value functions serves to synthesize the two attribute-specific values, albeit with different weights. The agent architecture must make use of a value function so that each candidate can be given an overall evaluation, and in MSIM, the agent architecture selects which value function to use on the basis of the current communicative context.

Chapter 7

Simulations of Multimodal Strategy Selection Using MSIM

7.1 Overview

In this chapter, we use MSIM to demonstrate computationally the impact of different VOCA interfaces on a communicator's mode strategies. Two types of impacts will be demonstrated. The first is *local* impact — that is, the impact on the specific strategy of using the mode of synthesized speech in isolation (the so-called *aided-unimodal* mode strategy, since it is afforded by the communication aid). The second is *global* impact — that is, the impact on the communicator's overall repertoire of mode strategies (which includes, in addition to the just-mentioned *aided-isolated* mode strategy, the *unaided* strategies and the *joint aided-unaided* strategies).

The results from three different simulation conditions — *unimodal interface VOCA*, *multimodal interface 1 VOCA*, and *multimodal interface 2 VOCA* conditions — will be described. The *interface* of the VOCA is the difference between the first two conditions: in the first it is unimodal, and in the second it is multimodal. The characteristics of the mode of synthesized speech are the same. The *quality* of the mode of synthesized speech is the difference between the second and third conditions: in the first it is the same as the one afforded by the unimodal interface, and in the second it is improved (faster and more intelligible). Thus, the third condition implements *bottleneck reduction*. The bottleneck reduction hypothesis, discussed previously in section 3.3.8, is that by increasing the information bandwidth of the VOCA interface (through the use of a multimodal interface), the mode of synthesized speech can be improved. (We make use of the middle condition simply in order to isolate effects due to increased mode conflict from effects of an improved mode of synthesized speech).

The parameter values for MSIM and results for each of these conditions will be described. The results will demonstrate that adding multimodality to the interface of a VOCA results in local, but not global, improvement. Thus, MSIM is a useful tool for demonstrating that a multimodal VOCA can afford a mode of synthesized speech that is improved over the one that is afforded by a unimodal VOCA, but the overall repertoire of mode strategies that it affords can be negatively affected.

7.2 The Simulation Conditions

We used MSIM to demonstrate the effect of the interface of a VOCA on a communicator's repertoire of mode strategies. In order to do this, we defined a set of three different simulation conditions, which is summarized in table 7.1.

Table 7.1 An overview of the simulation conditions invoked under MSIM.

Condition Name	Agent C's Mode Repertoire	Interference Set
unimodal interface VOCA (UI-VOCA)	gesture, vocalization, synthesized speech	VOCA conflicts with gesture
multimodal interface 1 VOCA (MM1-VOCA)	gesture, vocalization, synthesized speech	VOCA conflicts with gesture **and** vocalization
multimodal interface 2 VOCA (MM2-VOCA)	gesture, vocalization, **improved** synthesized speech	VOCA conflicts with gesture and vocalization

Agent C in all three simulation conditions performs the *multimodal referential communication task*. As described in section 5.6, MSIM expects its input file to contain a specification of the the set of entities to which the communicative agents may potentially refer. For this and the other two simulation conditions, the same set of eight entities was specified; they have been labelled $e_1, e_2, e_3, e_4, e_5, e_6, e_7$, and e_8. Each entity is uniquely identified by three of six possible semantic primitives: one for size (p_1 = SMALL, p_2 = LARGE), one for colour (p_3 = RED, p_4 = BLUE), and one for shape (p_5 = SPHERE, p_6 = CUBE). A complete list is given in table 7.2. Thus, in order to perform the task, agent C must perform a multimodal communicative action that signals a set of three semantic primitives.

Table 7.2 A list of the potential referents defined for the simulation conditions.

Entity ID	Description	Semantic Primitives
e_1	large red sphere	p_2, p_3, p_5
e_2	small red sphere	p_1, p_3, p_5
e_3	large blue sphere	p_2, p_4, p_5
e_4	small blue sphere	p_1, p_4, p_5
e_5	large red cube	p_2, p_3, p_6
e_6	small red cube	p_1, p_3, p_6
e_7	large blue cube	p_2, p_4, p_6
e_8	small blue cube	p_1, p_4, p_6

For each semantic primitive, one mode-specific sub-action was defined for each of the modes: gesture, vocalization, and synthesized speech. (Six semantic primitives by three modes yields 18 mode-specific sub-actions.) The mode-specific sub-actions for the SIZE semantic primitives were these: metaphoric gestures (e.g., the gesture of hands spread far apart for LARGE, the gesture of

hands drawing together for SMALL), vocalizations (e.g., dysarthric articulations of the words *large* and *small*), and the sequence of input actions to the VOCA that produces the synthesized speech (e.g., the words *large* and *small*, as produced by the text-to-speech module (TTS) of a VOCA once the corresponding input actions have been made). The sub-actions for the colour and shape primitives were defined similarly (e.g., the vocalization- and aided-mode-specific actions were defined analogously to the size primitives, and for the gesture-specific actions, instead of metaphoric gestures, pointing gestures were defined for the colour-related primitives, and iconic gestures were defined for the shape-related primitives).

In addition, the ordering relation $\{p_1, p_2\} < \{p_3, p_4\} < \{p_5, p_6\}$ was specified, meaning that all of the sub-actions for p_1 or p_2 (whichever the case may be) must be completed before those of other semantic primitives can be performed, and that the sub-actions for p_3 or p_4 (whichever the case may be) must be completed before those of the last semantic primitive can be performed. Note that with this condition, the only way for modes to be used simultaneously is if they are being used to signal mode-specific sub-actions *for the same semantic primitive*.

***Unimodal interface VOCA* condition** Agent C in the *unimodal interface VOCA* simulation condition characterizes an individual with a communication disorder. The mode of speech, although affected by dysarthria, is still available for the production of vocalizations. Gesture is available, although the motor movements are imprecise. The mode of synthesized speech is available through the use of a VOCA. The VOCA in this condition is defined to have an interface that requires key presses, which are *unimodal* input actions — hence the name *unimodal interface VOCA* (or *UI VOCA*, for short). We use MSIM to demonstrate the repertoire of mode strategies that is afforded to this individual.

For this specific joint activity, the VOCA requires a vocabulary of only six lexical items (two size adjectives, two colour adjectives and two shape nouns). We define the interface of the VOCA to be such that a lexical item is selected in two steps: first, the category of the lexical item is selected (e.g., one of *size adjective*, *colour adjectve*, or *shape noun*). We assume that each of these categories is associated with a single button on a touch-screen (the access technique is assumed to be direct). The VOCA's display dynamically updates so that the lexical items within the category are then presented, each of which is associated with a single button on the touch-screen. Thus, two direct-selection input actions are required.[1,2] We assume that once a lexical item has been selected, it is

[1]With such a small vocabulary, the VOCA interface could simply associate one button with each lexical item. However, this approach cannot be used for non-trivial vocabularies (which must be organized into hierarchies due to their large size). The VOCA interface in this condition is meant to be representative of these non-trivial VOCAs and organizes the six lexical items into a two-level hierarchy.

[2]The VOCA could exploit the assumption that has been made about the specific joint activity — the order of lexical categories is known in advance. The VOCA's interface could present the user with a sequence of three binary choices: the first for the size adjective to be used, the second for the colour adjective, and the third for the shape. For the sake of future scalability to other, more-complex joint activities, this technique will not be used.

passed to the VOCA's TTS module.[3,4]

Recall from section 6.5 that the state-transition model makes use of a number of different functions, each of which must be provided to MSIM. One of these functions is the *mode repertoire cost* function C_1. We assume the effort required to produce the input actions to the VOCA to be higher than the effort required to produce vocalizations or gestures. Thus, we define the cost of the aided mode to be relatively expensive (nine-tenths of the maximum cost), and the cost of the unaided modes of vocalization and gesture to be relatively cheap (one-fourth of the maximum cost; we assume they have the same cost). Thus, the values of the function C_1 are:

$$C_1(\text{UI-VOCA}) = 0.90$$
$$C_1(\text{VOCALIZATION}) = C_1(\text{GESTURE}) = 0.25$$

Another function that must be defined is the *mode repertoire interpretability difficulty* function R_1, which characterizes the likelihood that the communication partner will interpret a mode-specific sub-action as intended. In this condition, we define the unaided modes to be difficult to interpret — about three-quarters of the communication partners will misunderstand (i.e., a high index of difficulty), and the aided mode to be less likely to be misinterpreted — about one-quarter of the communication partners will misunderstand[5] (i.e., a low index of interpretability). Thus, the corresponding values of the function R_1 are:

$$R_1(\text{UI-VOCA}) = 0.25$$
$$R_1(\text{VOCALIZATION}) = R_1(\text{GESTURE}) = 0.75$$

***Multimodal interface 1 VOCA* condition** In this condition, we define the interface to the VOCA to be one that requires the communicator to make use of two modes of input. To select a lexical item, the user must (1) press the touch-screen button associated with the category of the lexical item, and (2) produce vocalization in order to select a particular item from within the category.[6] Treviranus et al. [1991] implemented, and demonstrated the utility of, precisely this sort of multimodal interface. Other than this difference, agent C in defined the same way as in the *UI-VOCA* condition.

This condition will be refered to as the *MM1 VOCA*, for short (and the results will be contrasted with those from the *MM2 VOCA* condition, to be described below).

With this interface, the use of synthesized speech conflicts with the mode of vocalization *and*

[3] An alternative would be for the VOCA to wait until three lexical items have been selected before passing the constructed string to the TTS module. However, an aided communicator using this type of VOCA would be forced to either use the mode of synthesized speech for *none* of the semantic primitives or *all* of them (with no option in between). In other joint activities, it might not be possible to signal every semantic primitive using a single word; in this case, the VOCA will need to implement a more sophisticated mechanism to determine when the constructed string should be passed to the TTS module. In some currently-available VOCAs, the constructed string is passed to the TTS module only when the user explicitly signals it.

[4] The VOCA might implement a scheme to prevent inconsistent utterances using synthesized speech (e.g., such as that corresponding to ill-formed sequences of lexical items, such as a sequence of two different size adjectives); if and when such errors are detected, even further input actions would be required by the aided communicator.

[5] This index of difficulty is higher than what was described in section 6.5.4 (e.g., synthesized speech interpretability was described at about 90%), but has been defined this way in order to emphasize the benefits of the multimodal VOCA.

[6] This requires that the user be able to reliably produce a repertoire of three distinct vocalizations and that the VOCA be trained to recognize them.

gesture, whereas in the *UI VOCA* condition (with the unimodal interface), the mode of synthesized speech conflicts *only* with the mode of gesture. In this condition, if the mode of synthesized speech is not used, then the mode of gesture doesn't conflict with the mode of vocalization. The qualities of the mode of synthesized speech are defined to be the same as the unimodal interface VOCA condition — for both conditions, the same values for the functions C_1 and R_1 are used; only the interference set that characterizes mode conflict changes.[7]

***Multimodal interface 2 VOCA* condition** Agent C in the *multimodal interface 2 VOCA* condition is defined the same way as in the *UI VOCA* and *MM1 VOCA* conditions; the only difference is the interface to the VOCA. In this condition, we define a hypothetical interface to the VOCA that allows the communicator to make use of multiple modes of input *simultaneously*. We will use the abbreviation *MM2 VOCA* for *multimodal interface 2 VOCA* condition henceforth.

The interface to this hypothesized VOCA presents to the user all six lexical items, arranged in two groups of three items each. The position of the lexical items within the group corresponds to the item's category (e.g., assuming a vertical arrangement, the *size* lexical items occupy the top position, the *colour* ones occupy the middle, and the *shape* nouns occupy the bottom position). Thus, each group consists of one lexical item from each of the three categories. One touch button is associated with each group. To select a lexical item, the user produces a vocalization and a key press — the vocalization selects the category of the lexical item (i.e., whether the top, middle, or bottom element is the target within a group) and the key press selects the item from within the category.[8] The interface of this hypothetical device would be multimodal and would be able to recognize the simultaneous use of two different modes.

A VOCA with this multimodal interface, in principle, could be used to produce synthesized speech in a shorter amount of time than one with a unimodal interface, since the two selections could be performed in parallel. Thus, in exchange for the increased mode conflict, this VOCA offers some sort of additional benefits beyond what can be obtained using a VOCA with a unimodal interface. Thus, this condition implements *bottleneck reduction*.

How are the benefits of bottleneck reduction manifested? First, the amount of time required to produce the synthesized speech is shortened.[9] Thus, the *MM2 VOCA* condition is different from the others in that the duration of time that is required to perform a sub-action using synthesized speech is reduced. However, the model for cost described in section 6.5.3 does not take into account the time durations of the mode-specific sub-actions that compose a multimodal communicative action. In MSIM, a decrease in the amount of time required to produce synthesized speech *does* have an

[7] One could argue that the cost value of C_1(MM1-VOCA) should be different than C_1(UI-VOCA), in order to take into account the difference between the cost of producing the vocalization compared to the cost of producing the input gesture. However, we have assumed that the costs are the same and, thus, the same value as C_1(UI-VOCA) has been kept (0.90). This allows us to isolate the effect of increased mode conflict.

[8] This requires that the user be able to reliably produce a repertoire of three distinct vocalizations and that the VOCA be trained to recognize them. The button press and the vocalization can be done simultaneously or in sequence, in either order.

[9] In MSIM, this is implemented by manipulating the values passed to the Δ-criterion. Recall from section 6.3 that the minimum and maximum timesteps for each of the mode-specific sub-actions must be specified to the agent architecture. These values are parameters to the Δ criterion, which, in part, determines the set of candidate surface realizations. For the *UI VOCA* and *MM1 VOCA* conditions, the minimum and maximum timesteps are defined to be two units of time (i.e., $\Delta_{M_{\text{UI-VOCA}}} = \{a_{\text{UI-VOCA},i}, 2, 2\}$, $\forall i = 1, \ldots, 6$);$\Delta_{M_{\text{MM1-VOCA}}} = \{a_{\text{MM1-VOCA},i}, 2, 2\}$, $\forall i = 1, \ldots, 6$). For the *MM2 VOCA* condition, the minimum and maximum timesteps are defined to be one unit of time (i.e., $\Delta_{M_{\text{MM2-VOCA}}} = \{a_{\text{MM2-VOCA},i}, 1, 1\}$, $\forall i = 1, \ldots, 6$). (The mode-specific sub-actions for the unaided modes were defined to be one unit of time each in all conditions.).

effect on the types of candidates that are generated (they require fewer timesteps), but the cost model is unfortunately not sensitive to this. This is a shortcoming of the way in which MSIM calculates the physical cost of performing a multimodal communicative action (see section 8.3.1 for a thorough discussion of avenues for its improvement).

To compensate for this shortcoming, we defined the mode of synthesized speech to be more interpretable than in the *MM1 VOCA* condition. This improvement could be achieved through the use of prosodic emphasis in the synthesized speech produced by the TTS module. Before the constructed string is passed to the TTS module, the hypothetical VOCA will identify and annotate the lexical item (or items) that contrast from the previous referent.[10] The use of different prosodic contours could improve the interpretability of the synthesized speech and the success of the task outcome. To simulate this, in the *MM2 VOCA* condition, the interpretability index of the mode of synthesized speech was set at its limit so that the communication partner never misunderstands the mode of synthesized speech:

$$R_1(\text{MM2-VOCA}) = 0.00$$

This function value improves the mode of synthesized speech from the *MM1 VOCA* condition, where it was defined to be $R_1(\text{UI-VOCA}) = 0.25$, and is the reward for the additional expense of the multimodal input actions.

Scaling Factors Recall from sections 5.6 and 6.5 that MSIM's models of the cost and of the interpretability of a multimodal communicative action allows for primitive-specific scaling factors. However, for these conditions, we choose not to investigate the possible effects arising from these scaling factors and no primitive-specific scaling was specified. For all conditions, the *semantic primitive interpretability scaling* function R_2 and the *semantic primitive cost scaling* function C_2 were defined as follows:[11]

$$\forall p_i \in \mathcal{X} \text{ and } \forall M_j \in \mathcal{M}, \quad C_2(p_i, M_j) = R_2(p_i, M_j) = 1.00$$

Number of Iterations Another parameter of MSIM is the value k for each simulation condition, which is the number of times simulations should be invoked for the given condition (see section 5.6). The value $k = 10$ was used for all of the simulation conditions. However, this value is of little consequence in the subsequent discussions, since we will focus on the composition of the set of candidates (from one invocation under a given condition), rather than the actual candidates that MSIM selected over a number of repeated invocations. (Recall from section 5.6 that the behaviour of agent C is not entirely deterministic; in the case of ties among the top-ranked candidates, agent C choses from among them randomly. The value of k in effect determines the size of the sample of this tie-breaking behaviour.)

[10] For instance, if the current intended referent differs from the previous one only with respect to colour, then the colour adjective would be emphasized (e.g., if the intended referent in the preceeding task was the large blue cube, then the emphasis in the utterance "the large RED cube" signals that the difference in colour is salient). Emphasis of all three lexical items is not possible (e.g., if the size, colour, and shape of the intended referent are all different from those of the previous referent).

[11] The set $\mathcal{X}$ is the set of all semantic primitives known to the interlocutor. See p. 71 for the defintion.

State Transition Model Recall from section 6.5 that two different state transition models were developed, g_1 and g_2. In the invocations of MSIM described here, the latter state transition model was used (since the former does not take into account any possible additional costs or benefits of using modes with redundancy or simultaneity).

Equivalence Relation Our analysis of MSIM's output will focus on the characteristics of the sets of candidate multimodal surface realizations that the chooser agent C derives for its communication plan (which has the form *plan-ref*(e_i), where e_i represents the randomly-selected referent). As described in section 6.4.2, MSIM uses an equivalence relation to partition the set of candidate surface realizations into equivalence classes, where each equivalence class corresponds to a distinct mode strategy. In simulations described here, MSIM uses the equivalence relation $\dddot{\sim}$, which partitions the candidate set into 14 equivalence classes.[12] The surface realizations in a given equivalence class each have two *goal-specific values* (one with respect to the domain goal and one with respect to the procedural goal). The set of domain-goal-specific values and the set of procedural-goal-specific values of the surface realizations in an equivalence class will be used to characterize the *goal-specific values* of the mode strategy that corresponds to that equivalence class.

Communication Partner Familiarity, Tolerance to Fatigue The surface realizations in a given equivalence class each have an *overall value*. The set of overall values of the surface realizations in an equivalence class will be used to characterize the *overall value* of the mode strategy that corresponds to that equivalence class. However, recall that the derivation of a candidate's overall value depends on the particular value function that is used (which determines the relative weights of these two goal-specific values). As described in 6.5.5, MSIM chooses one of six possible value functions to use, on the basis of the characteristics of the communication scenario (specifically, the familiarity of the communication partner and the communicator's own tolerance to fatigue). In the discussions below, the results for all six value functions will be described.

7.3 Discussion of the *unimodal VOCA* condition

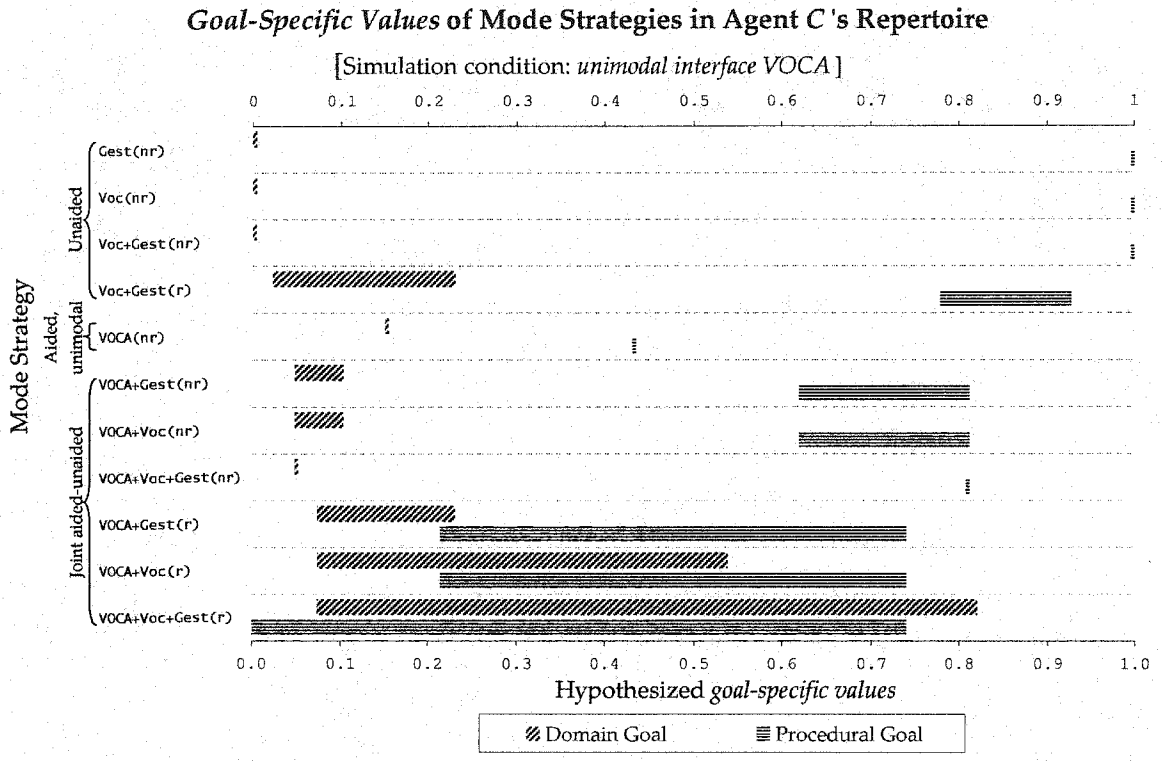
We first describe MSIM's output for the *UI VOCA* condition in order to introduce and to discuss the graphical presentations of the simulation results.

Figure 7.1 provides a summary of the *goal-specific values* for each of the mode strategies in agent C's repertoire. The y-axis in figure 7.1 provides a summary of the mode strategies that are available to agent C in the *UI VOCA* condition; they have been categorized as *unaided* (i.e., the strategy does not entail the use of the VOCA), *aided-unimodal* (i.e., the strategy entails the use of the mode of synthesized speech that is afforded by the VOCA in isolation), and *joint aided-unaided* (i.e., the strategy entails the combined use of the mode of synthesized speech and the other modes, which do not rely on the VOCA). (See section 4.3 for the description of these types.)

Agent C's repertoire consists of 11 mode strategies, rather than the expected repertoire of 14. As described in the previous section, MSIM makes use of the equivalence relation

[12]The equivalence relation $\sim$ partitions the set of candidates too coarsely (7 equivalence classes), whereas the relation $\approx$ partitions it too finely (28 equivalence classes). See section 6.4.2 for further detail.

Figure 7.1 Goal-specific values for the mode strategies in the *unimodal VOCA* condition. The suffix (r) indicates a redundant mode strategy (i.e., it has at least one semantic primitive that is multiply-signalled), whereas the suffix (nr) indicates a non-redundant mode strategy.



$\ddot{\sim}$, which partitions the set of all candidate multimodal surface realizations into 14 equivalence classes. Each equivalence class represents a mode strategy and contains all of the multimodal surface realizations that make use of that particular mode strategy. In particular, there is a *redundant* and a *non-redundant* variant for each of seven mode strategies: three *unaided* strategies (the use of the mode of gesture in isolation, the use of the mode of vocalization in isolation, and the joint use of gesture and vocalization: Gest(nr)/Gest(r), Voc(nr)/Voc(r), and Voc+Gest(nr)/Voc+Gest(nr), respectively); one *aided-unimodal* mode strategy (the use of the mode of synthesized speech in isolation: VOCA(nr)/VOCA(r)); and three *joint aided-unaided* mode strategies (the use of the mode of synthesized speech accompanied by the mode of gesture, vocalization, or both: VOCA+Gest(nr)/VOCA+Gest(r), VOCA+Voc(nr)/VOCA+Voc(r), VOCA+Voc+Gest(nr)/VOCA+Voc+Gest(r), respectively). A redundant mode strategy is one in which a semantic primitive is signalled by multiple mode-specific sub-actions (i.e., it is multiply-signalled); such strategies are indicated by the suffix (r).

In the *UI VOCA* simulation condition (as well as all of the others described in this chapter), for each semantic primitive, only one mode-specific sub-action has been defined for each mode. So if a semantic primitive is to be multiply-signalled, then the multiple sub-actions must be performed using different modes. Thus, it is not possible for the aided-unimodal mode strategy to be redun-

dant, nor is it possible for the unaided mode strategies of using gesture or vocalization in isolation to be redundant. Thus, the three equivalence classes of unimodal redundant mode strategies (i.e., VOCA(r), Gest(r), and Voc(r)) do not occur (which is why agent C has a repertoire of 11 mode strategies instead of 14 in these simulation conditions).

The x-axis of figure 7.1 corresponds to the *goal-specific values* for the mode strategies (as opposed to the *overall values* for the mode strategies; to be discussed below). The mode strategies correspond to equivalence classes; each equivalence class, in turn, is characterized by the goal-specific values of the surface realizations contained within. Thus, a mode strategy can be characterized by one or more sets of values (in the figures here, for the sake of simplicity, we show only the ranges of values, rather than the distributions over the values). In figure 7.1, for each mode strategy (equivalence class), two sets of values are shown: one corresponds to the goal-specific values with respect to the domain goal, and the other to the procedural goal-specific values. The values were derived on the basis of the attributes of the consequence states that were specified by the state-transition model that was described in section 6.3. Each set of values is represented by one horizontal bar in the graph, which illustrates the minima and maxima of the goal-specific values for the corresponding equivalence set. The shortest horizontal bars correspond to equivalence classes whose extrema are equal.

The AAC research literature notes that the *unaided* mode strategies require less effort than the aided mode of synthesized speech, but that they are less interpretable by communication partners [Blischak and Lloyd, 1996; Garrett and Beukelman, 1998]. This is demonstrated in figure 7.1 by the relatively high evaluations of the unaided-unimodal strategies (i.e., gesture or vocalization in isolation) with respect to the procedural goal and by their relatively low evaluations with respect to the domain goal. For the unaided mode strategies, the severity of these effects is mitigated if the strategy is multimodal and make use of redundancy — compare the unaided-multimodal strategy of using gesture and vocalization with redundancy to the unaided-unimodal strategies (i.e., compare the horizontal bars of Voc+Gest(r) to those of Gest(nr) and Voc(nr)). The aided mode of synthesized speech requires more effort (reflected in its much lower evaluation with respect to the procedural goal than the unaided mode strategies), but can be more readily understood by the communication partner (reflected in its much higher evaluation with respect to the domain goal than the unaided mode strategies). The horizontal bars of the multimodal strategies are much larger, since there is a larger range of values (which corresponds to the larger number of ways to combine mode-specific sub-actions when multiple modes can be used); the figure exhibits the pattern that the more modes a strategy uses, the more physical effort it entails but the greater its interpretability.

The *overall value* of a mode strategy is a weighted sum of the two goal-specific values, where those weightings depend on the value function that is used. MSIM tailors the value function to the communicative context, which is characterized by two parameters of the simulation condition: the *familiarity* of the communication partner and the communicator's *tolerance to fatigue*. The rationale for the mapping from the parameter space to the different value functions was described in section 6.5.5. In order to compare two mode strategies, we compare the *overall values* of one mode strategy to the *overall values* of the other. The weightings used by each of the six value functions are summarized in table 7.3.

Figure 7.2 shows the overall values of the mode strategies in agent C's repertoire in the *UI*

Table 7.3 An overview of the weightings used by each of the six value functions (see section 6.5.5 for further discussion).

Simulation Condition communication partner familiarity, tolerance to fatigue	**Value Function Used**	**Weight of Domain Goal**	**Weight of Procedural Goal**
low, high	V_1	1.000	0.000
low, med	V_2	0.750	0.250
low, low	V_3	0.500	0.500
med, high	V_3	0.500	0.500
med, med	V_4	0.375	0.625
med, low	V_5	0.250	0.750
high, low	V_6	0.000	1.000
high, med	V_6	0.000	1.000
high, high	V_6	0.000	1.000

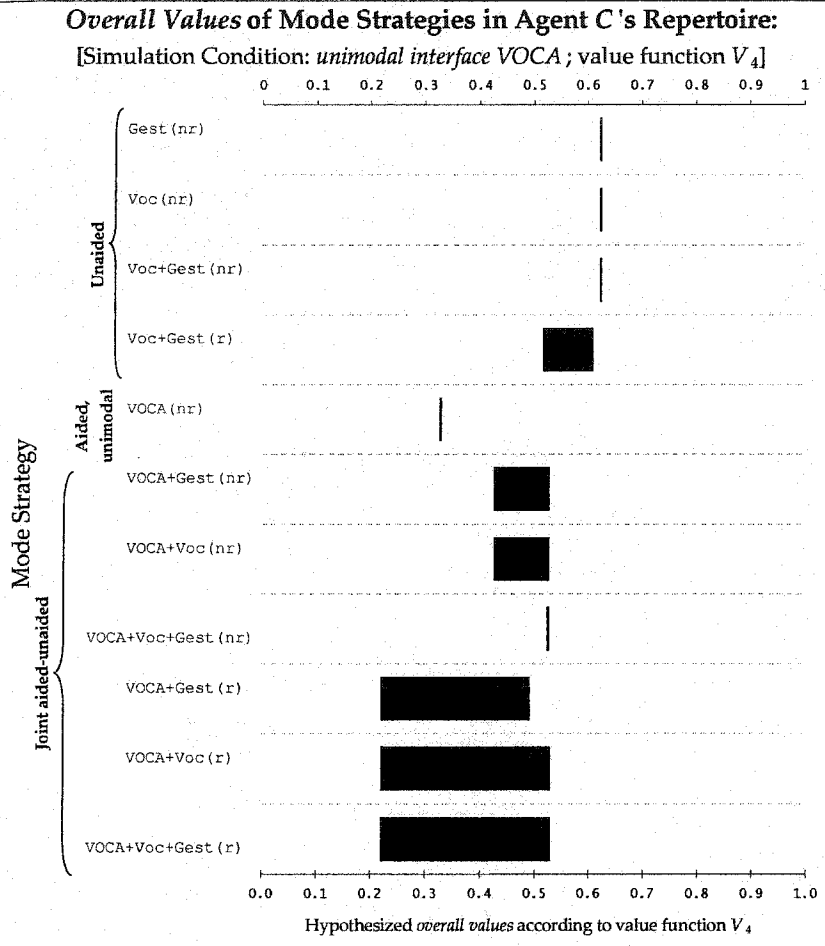
VOCA condition when the value function V_4 is used by MSIM. The *overall values* of a mode strategy are referred to as the mode strategy's profile. **The relative position of the horizontal bars for each of the mode strategies serves to rank the mode strategies in a given communication scenario.** The bottom diagram of figure 7.2 shows the ranking of the mode strategies in agent C's repertoire (the maximum values of the horizontal bars were used to derive the ranking).[13]

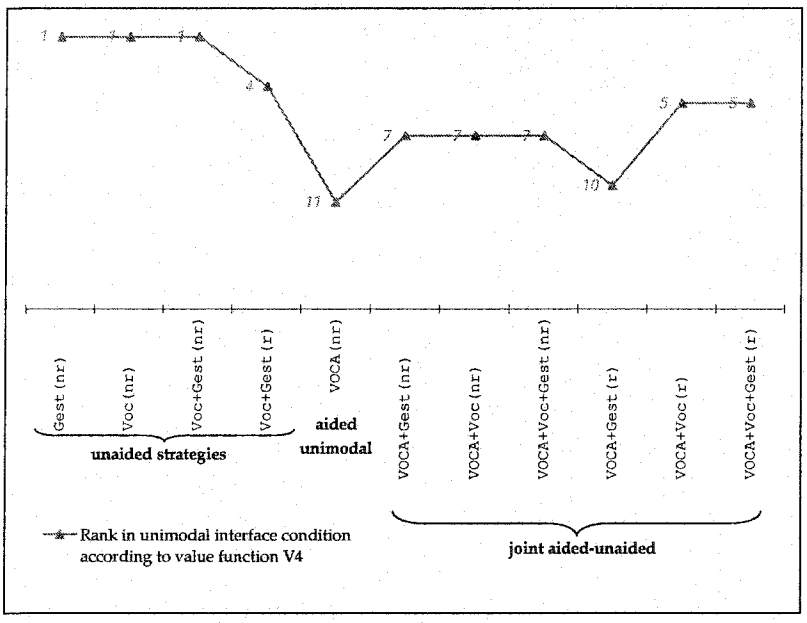
The results in figure 7.2 demonstrate that, if the communication partner is moderately familiar and if the communicator's tolerance to fatigue is moderate, the unaided mode strategies are best, followed by the joint aided-unaided mode strategies (with redundancy). In these situations, the least optimal mode strategy to use is the aided-unimodal mode strategy. MSIM demonstrates that, in these situations, the benefits of the VOCA are outweighed by its cost. The profiles that are derived when the other five value functions are used can be found in appendix A.5.

The results of the simulation, in general, are that the unaided mode strategies are optimal in cases in which the communication partner is not unfamiliar (e.g., when value functions V_3, V_4, V_5, and V_6 are used) or when the communication partner is familiar but the communicator doesn't have a high tolerance for fatigue. In cases in which the communication partner is unfamiliar (and tolerance to fatigue is not low — that is, when the value functions V_1 and V_2 are used), the high cost of the VOCA is outweighed by its benefits, and its use becomes warranted. Even so, MSIM shows that the use of the VOCA in combination with other modes (i.e., joint aided-unaided mode strategies) is always ranked higher than the use of the VOCA in isolation (i.e., the aided-unimodal strategy). Only if the communicator's tolerance to fatigue is high (e.g., when the value function V_1 is used) does the aided-unimodal strategy become relatively highly ranked (but even in this case, the unaided multimodal strategy with redundancy is ranked higher). These patterns can be seen in the rankings found in appendix A.5.

[13]The rank of an element is defined here to be its position if the elements were to be sorted (after the duplicates have been removed).

Figure 7.2 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* conditions when the value function V_4 is used (top graph) and their ranking, using the maximum overall value of each mode strategy (bottom diagram).





7.4 Comparing the *VOCA* Conditions

In this section, we will contrast the output for the *unimodal interface VOCA* and the *MM2 VOCA*, in order to see MSIM's demonstration of the consequences of bottleneck reduction on agent *C*'s repertoire of mode strategies. As described in section 7.2, the latter condition implements the *bottleneck reduction*. First, we will contrast the output for the *UI VOCA* and the *MM1 VOCA*, in order to see MSIM's demonstration of the consequences of increased mode conflict on agent *C*'s repertoire of mode strategies. Then we will contrast the output from the *MM1 VOCA* and *MM2 VOCA* conditions, in order to see MSIM's demonstration of the consequences of an improved mode of synthesized speech. Last, we will contrast the output from the *UI VOCA* and *MM2 VOCA* conditions, in order to see MSIM's demonstration of the consequences of the collective effect of increased mode conflict and an improved mode of synthesized speech.

The Effect of an Increase in Mode Conflict

In figure 7.3, the *goal-specific values* of each of the mode strategies for each of the three conditions are illustrated: the *UI VOCA* condition, the *MM1 VOCA* condition, and the *MM2 VOCA* condition. The simulation outputs for each of these three conditions are shown aligned vertically, in the order just given. MSIM produces *two* horizontal bars for each mode strategy for *each* of three conditions (thus, each mode strategy has six horizontal bars). Results that will be discussed specifically are labelled on the figure by circled numbers.

MSIM demonstrates that an *increase* in mode conflict has a detrimental effect on the joint aided-unaided mode strategies. The simulation outputs that demonstrate this are labelled (1) in figure 7.3. The detrimental effect is manifested in the joint aided-unaided mode strategy in which the mode of synthesized speech and vocalization are used together. This mode strategy is labeled VOCA+Voc(r) in the figure. Note in figure 7.3 that the domain-goal-specific values for this mode strategy change from the *UI VOCA* condition to the *MM1 VOCA* condition, as a consequence of increased mode conflict. (Specifically, the highest domain-goal-specific value of this mode strategy in the *UI VOCA* condition is just over half the maximum possible value, whereas in the *MM1 VOCA* condition, the largest value has been reduced by half. The smallest domain-goal-specific value remains the same.)

Through which mechanism is the domain-goal-specific value of the mode strategy VOCA+Voc(r) reduced? In short, the sets of candidate surface realizations that are generated by agent *C* differ between the two conditions, and different candidate sets yield different sets of goal-specific values. A candidate surface realization in which the modes of vocalization and synthesized speech are used simultaneously is legitimate in the *UI VOCA* condition — in this condition, the modes of synthesized speech and vocalization do not conflict. Such a candidate, however, is not legitimate in the *MM1 VOCA* condition because the modes of synthesized speech and vocalization are conflicting; such a candidate violates the *mode conflict avoidance criterion* and MSIM disallows it (see section 6.3.2). Thus, surface realizations that are legitimate candidates in the *unimodal VOCA* condition are not legitimate in the *MM1 VOCA* condition. The equivalence class that corresponds to the mode strategy VOCA+Voc(r) in the unimodal condition differs from the corresponding equivalence class in the multimodal condition because a number of surface realizations have been eliminated from consideration.

Table 7.4 Characteristics of the set of candidates generated in the *UI VOCA* and the *MM1 VOCA* simulation conditions.

(a) Characteristics of the equivalence classes produced in the *UI VOCA* condition.

	Redundancy (nr)	Redundancy (r)		
Gest.	1 (0.01%)	0 (0.00%)	1 (0.01%)	13824 (100.00%)
Voc.	1 (0.01%)	0 (0.00%)	1 (0.01%)	
Voc.+Gest.	6 (0.04%)	117 (0.85%)	123 (0.89%)	
VOCA	1 (0.01%)	0 (0.00%)	1 (0.01%)	
VOCA+Gest.	6 (0.04%)	56 (0.41%)	62 (0.45%)	
VOCA+Voc.	6 (0.04%)	208 (1.50%)	214 (1.55%)	
VOCA+Voc.+Gest.	6 (0.04%)	13416 (97.05%)	13422 (97.09%)	
	27 (0.20%)	13797 (99.80%)		
	13824 (100.00%)			

(b) Characteristics of the equivalence classes produced in the *MM1 VOCA* condition.

	Redundancy (nr)	Redundancy (r)		
Gest.	1 (0.02%)	0 (0.00%)	1 (0.02%)	5832 (100.00%)
Voc.	1 (0.02%)	0 (0.00%)	1 (0.02%)	
Voc.+Gest.	6 (0.10%)	117 (2.01%)	123 (2.11%)	
VOCA	1 (0.02%)	0 (0.00%)	1 (0.02%)	
VOCA+Gest.	6 (0.10%)	56 (0.96%)	62 (1.06%)	
VOCA+Voc.	6 (0.10%)	56 (0.96%)	62 (1.06%)	
VOCA+Voc.+Gest.	6 (0.10%)	5576 (95.61%)	5582 (95.71%)	
	27 (0.46%)	5805 (99.54%)		
	5832 (100.00%)			

The elimination of these candidate surface realizations is the reason that the VOCA+Voc(r) equivalence class that is derived in the *MM1 VOCA* condition is smaller than the corresponding class that is derived in the *UI VOCA* condition. The sizes of all of the equivalence classes are given in table 7.4. These tables itemize the number of surface realizations that make use of each type of mode strategy. Beside each mode strategy count, the percentage with respect to the whole set is given. (The sum of the percentages may be slightly incorrect due to rounding error.) The equivalence class corresponding to the mode strategy VOCA+Voc+Gest(r) is also changed between the two conditions; this will be discussed further below.

Note that from the unimodal condition to the multimodal condition, the domain-goal-specific values of the VOCA+Voc(r) mode strategy are affected, but not the procedural-goal-specific values. The reason is that the surface realizations that are excluded by agent C in the latter condition (but not the former) are those that have relatively high values with respect to the domain goal. The left-hand side of figure 7.4 shows a scatterplot of the candidates in the VOCA+Voc(r) equivalence class that is derived in each of the two conditions. We can see in the scatterplot that the candidates that do not get generated in the multimodal VOCA condition are precisely those that have the relatively higher domain-goal specific values. The model of interpretability, from which the domain-goal-specific values are derived, rewards for the synergistic effect of simultaneously-performed mode-specific sub-actions. But the high-simultaneity candidates are those that are adversely affected by the mode conflict in the multimodal condition. Thus, the domain-goal-specific values for the VOCA+Voc(r) mode strategy in the multimodal VOCA condition are lower than in the unimodal interface VOCA condition. The scatterplot also demonstrates that the extrema of the interval of

procedural-goal-specific values for this mode strategy do not change between the two conditions.

Table 7.4 also shows that the equivalence class of multimodal surface realizations that correspond to mode strategy VOCA+Voc+Gest(r) is different between the two conditions. However, we don't see any effect in the goal-specific values in figure 7.3. The right-hand side of figure 7.4 shows a scatterplot of the candidates in the VOCA+Voc+Gest(r) equivalence class that is derived in each of the two conditions. We see that the top- and bottom-ranking candidates with respect to each of the goals have not been eliminated from the equivalence class in the multimodal VOCA conditions. Thus, the extrema of the intervals for the mode strategy VOCA+Voc+Gest(r) in figure 7.3 are not changed between the two conditions.

The equivalence classes of multimodal surface realizations that correspond to the unaided strategies (i.e., Gest(nr), Voc(nr), Voc+Gest(nr), and Voc+Gest(r)) and the aided-unimodal strategy (i.e., VOCA(nr)) do not differ between the two conditions (table 7.4 shows the sizes of the equivalence classes are the same between the two conditions, and inspection of the output files shows that the members of the classes are the same too). For this reason, the horizontal bars corresponding to these mode strategies do not differ between the simulation outputs from the *UI VOCA* and the *MM1 VOCA* conditions in figure 7.3.

The Effect of Improving the Mode of Synthesized Speech

MSIM demonstrates that increasing the interpretability of the aided mode of synthesized speech has a beneficial effect on all of the mode strategies that make use of that mode: the aided-unimodal strategy (i.e., VOCA(nr)) and all the joint aided-unaided strategies. These improvements are demonstrated in the contrast between the simulation outputs from the *MM1 VOCA* and *multimodal interface VOCA* conditions and are labeled (2)–(5) in figure 7.3.

Combined Effect

The combined effects of an increase in mode conflict and an improved mode of synthesized speech are demonstrated in the contrast between the simulation outputs from the *UI VOCA* and *MM2 VOCA* conditions. Figure 7.3 demonstrates that:

- The domain-goal-specific values of the mode strategy of using the aided mode in isolation are *improved* by the modifications from the *UI VOCA* to the *MM2 VOCA* simulation condition (note (2) in figure 7.3).
- The domain-goal-specific values of the joint aided-unaided strategy VOCA+Voc(r) (i.e., the strategy of using vocalization and synthesized speech together) were *negatively* affected by mode conflict, which arises from the multimodal interface, but *positively* affected by benefits of the multimodal interface. However, the positive effects were not sufficient to outweigh the negative effects. (See notes (1) and (3) in figure 7.3)
- The domain-goal-specific values of the joint aided-unaided strategies other than VOCA+Voc(r) were *positively* affected by benefits of the multimodal interface (notes (4) and (5) in figure 7.3).

Consequences for the Bottleneck Reduction Hypothesis

We now contrast the *UI VOCA* and *MM2 VOCA* conditions in order to examine MSIM's demonstration of the effect of bottleneck reduction on agent *C*'s repertoire of mode strategies.

We begin by discussing the impact of bottleneck reduction on the mode strategy evaluations with respect to each of the value functions. For half of the value functions, bottleneck reduction resulted in very little or no change. In particular, the value function V_6 evaluates overall value solely on the basis of the procedural-goal-specific values of the mode strategies. The VOCA improvements affect the domain-specific-goal values specifically, and the value function V_6 does not take these into account. Thus, the evaluation of the overall values of the mode strategies do not change between the conditions. Similarly, the value functions V_4 and V_5 evaluate overall value primarily (but not exclusively) on the basis of the procedural-goal-specific values of the mode strategies (the domain-goal-specific values contribute very little). Thus, the VOCA improvements had little impact on the mode strategy evaluations. As shown in figures A.8, A.9, A.10, and A.11 in appendix A.5, the VOCA improvements either did not affect the mode strategy rankings or affected them by only a few positions (the paragraphs below discuss the format of these graphs in more detail).

In the value functions V_1, V_2, and V_3, the weights allocated to the domain goal are sufficiently high that the VOCA improvements have an effect on the evaluations of agent *C*'s repertoire of mode strategies from the *UI VOCA* to the *MM2 VOCA* conditions. We begin with value function V_2, since it produces the most drastic differences between the two conditions. When the value function V_2 is used, the overall values of the mode strategies are based primarily on the domain-goal-specific values, with some contribution from the procedural-goal-specific values (the weights are 0.75 and 0.25, respectively).

The overall values of the mode strategies in agent *C*'s repertoire are shown in figure 7.5, according to the value function V_2 for each the *UI VOCA* and *multimodal VOCA* conditions (the mode strategy values for the first condition are shown as black bars, whereas those for the latter are shown as grey bars). The figure shows that the unaided mode strategies are unaffected by the VOCA improvements, that the aided-unimodal strategy is improved, and that *most* of the joint aided-unaided mode strategies are improved. The exception is the VOCA+Voc(r) strategy, which has been adversely affected — the improvements to the VOCA improve this mode strategy to some degree, but those improvements are outweighed by the negative effects of mode conflict (this is shown more specifically in figure 7.3, labelled (1) and (3)).

We use the maximum value of each mode strategy in a condition to rank the mode strategies in agent *C*'s repertoire. Rankings for each of the two conditions are shown in figures 7.6 and 7.8 (for value function V_2 and V_1, respectively). In these figures, boxes have been used to highlight the results associated with the aided-unimodal strategy and two of the joint aided-unaided mode strategies. We will consider specifically these three mode strategies when assessing the global impact of bottleneck reduction.[14]

Figure 7.6 illustrates that, when value function V_2 is used, bottleneck reduction resulted in an

[14]The other joint aided-unaided strategies are not considered because they are non-redundant (and thus are unaffected by mode-conflict). The mode strategy VOCA+Voc+Gest(r) is not considered because it entails the use of three different modes (all the other joint aided-unaided strategies entail the use of two modes).

increase in overall value of the aided-unimodal strategy, VOCA(nr), a decrease in the overall value of the VOCA+Voc(r) mode strategy, and increases in the overall values of the other joint aided-unaided mode strategies. The mode strategy rankings were also affected. First, the improvement to the aided-unimodal strategy was enough to promote it from a rank of 11 in the *UI VOCA* condition (the worst ranking) to a rank of 5 in the *MM2 VOCA* condition. Second, the ranks of the joint aided-unaided mode strategies were almost all improved to some degree. One exception is the VOCA+Voc(r) strategy, whose value declined.[15] If we consider the rankings of the mode strategies, then bottleneck reduction resulted in global improvement (for VOCA(nr), VOCA+Gest(r), and VOCA+Voc(r), the rank changes are +6, +4, and −1). However, if we consider the overall values of the mode strategies, then bottleneck reduction resulted in global detriment (for VOCA(nr), VOCA+Gest(r), and VOCA+Voc(r), the overall value differences are +0.058, +0.043, and −0.173).

The overall values of the mode strategies according to the value function V_1 for both of the conditions are shown in figure 7.7. The effect of bottleneck reduction when value function V_1 is used is similar to that when value function V_2 is used.

[15]The other exception is the VOCA+Voc+Gest(nr) strategy, whose rank remained the same.

Figure 7.3 Goal-specific values for the mode strategies for three different conditions: the *UI VOCA* condition, the *MM1 VOCA* condition, and the *multimodal VOCA* condition. The suffix (r) indicates a redundant mode strategy (i.e., it has at least one semantic primitive that is multiply-signalled), whereas the suffix (nr) indicates a non-redundant mode strategy. The first data series in the chart below is the same data series that was shown in figure 7.1, but has been repeated for the sake of contrast. See the main text for explanation of the components that are labeled with circled numbers.

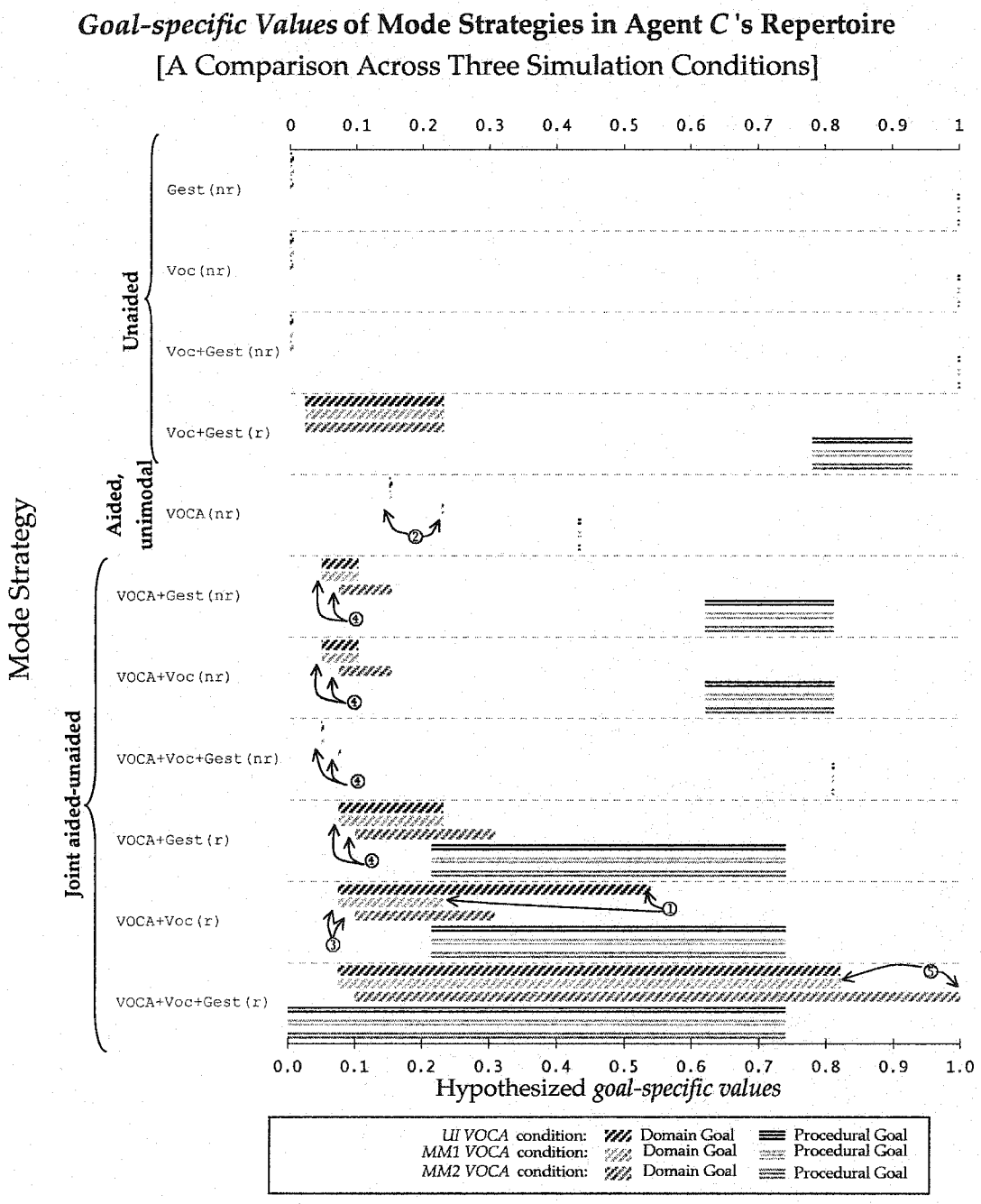


Figure 7.4 Goal-specific values for selected equivalence classes of candidate surface realizations in both the *UI VOCA* and *MM1 VOCA* simulation conditions.

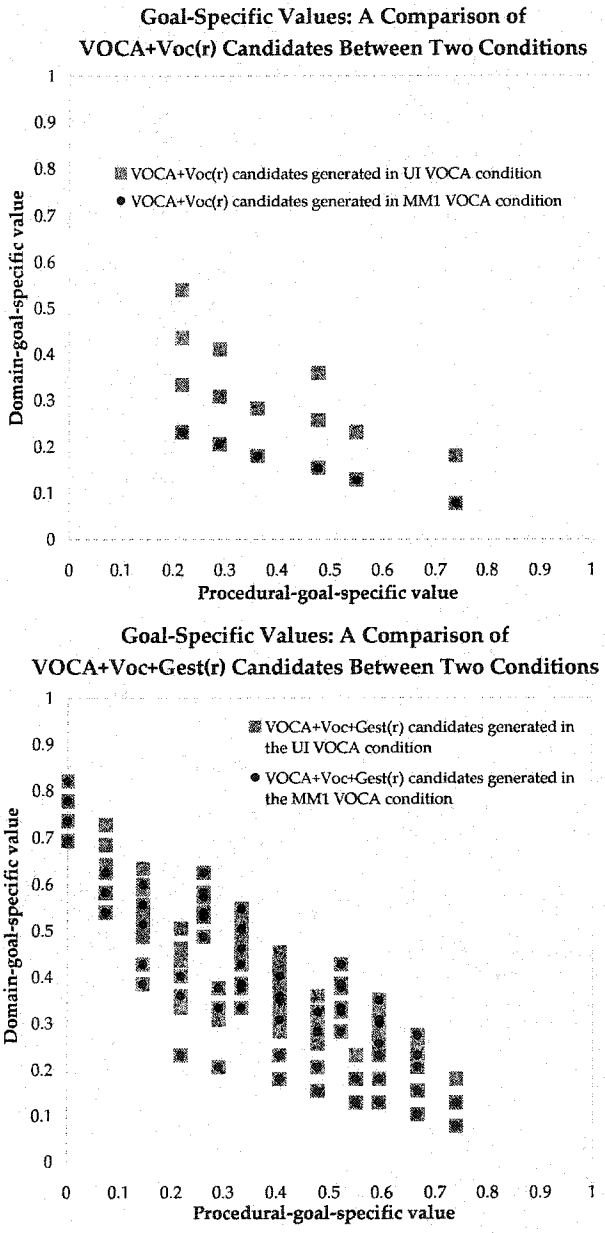


Figure 7.5 Mode strategy profiles for the *UI VOCA* and *MM2 VOCA* conditions when value function V_2 is used.

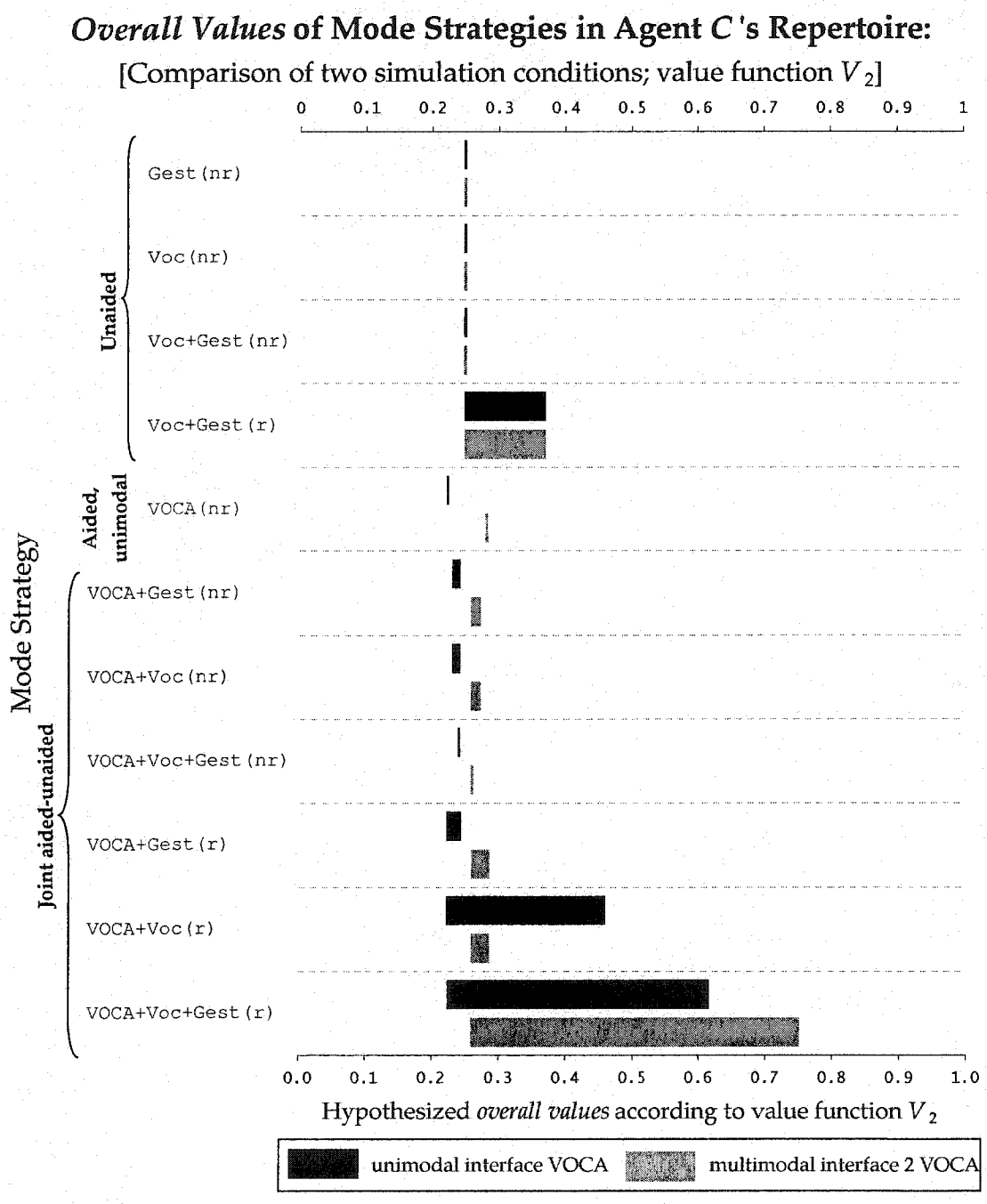


Figure 7.6 Rankings of the mode strategies in agent *C*'s repertoire under the *UI VOCA* and the *MM2 VOCA* conditions, when value function V_2 is used. For the mode strategies VOCA+Gest(r), and VOCA+Voc(r), the rank changes are +6 (from 11 to 5), +4 (from 7 to 3), and -1 (from 2 to 3).

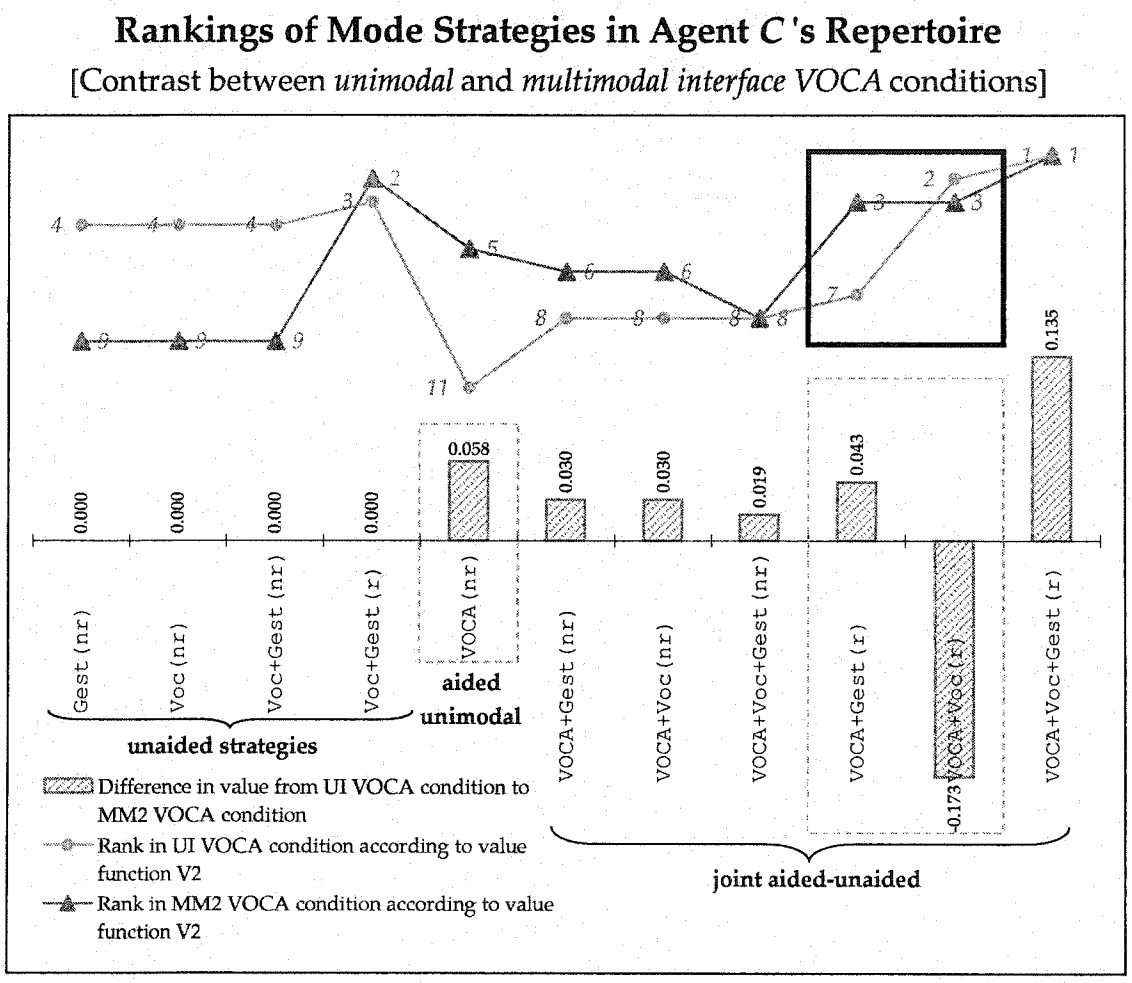


Figure 7.7 Mode strategy profiles for the *UI VOCA* and *MM2 VOCA* conditions when value function V_1 is used.

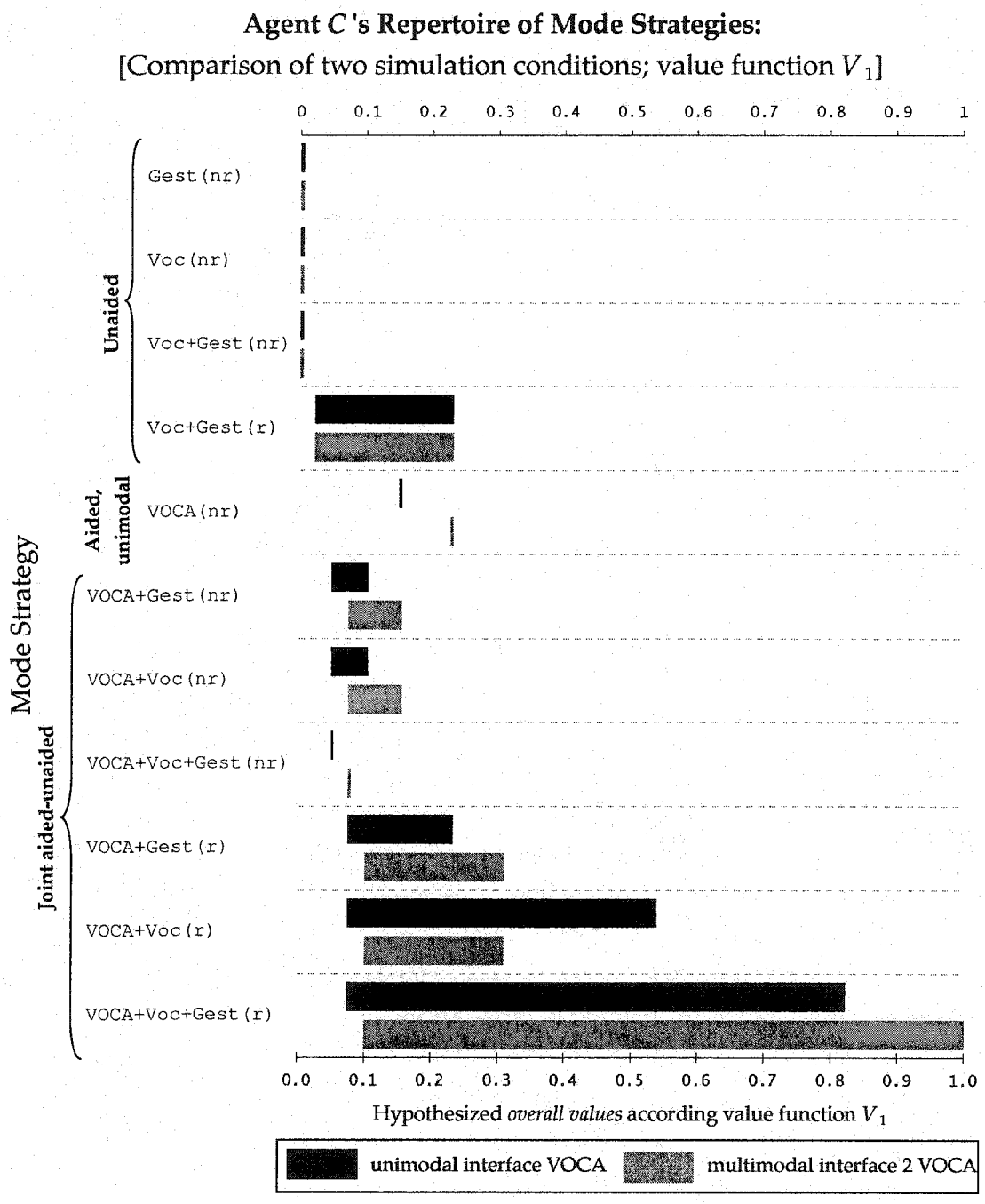
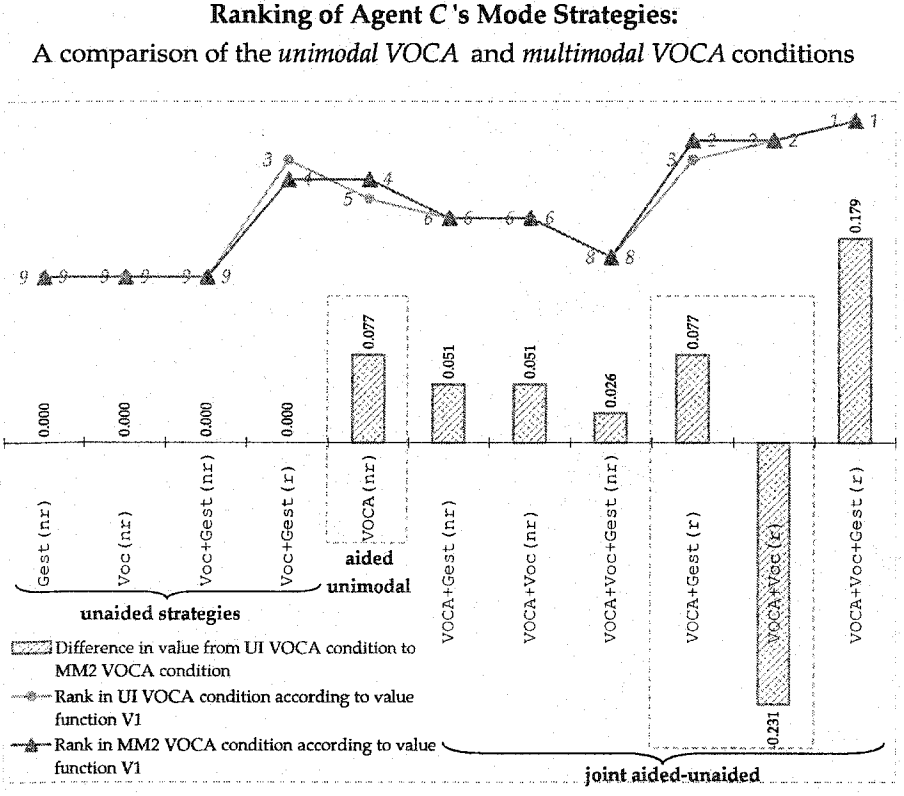


Figure 7.8 illustrates that, when value function V_1 is used, the mode strategies that are most highly ranked are those that entail redundancy, followed by the aided-unimodal strategy. Bottleneck reduction resulted in an increase in overall value of the aided-unimodal strategy, VOCA(nr), a decrease in the overall value of the VOCA+Voc(r) mode strategy, and increases in the overall values of the other joint aided-unaided mode strategies, the same pattern as when value function V_2 is used. However, with this value function, the ranking of the aided-unimodal strategy VOCA(nr) improves in the *MM2 VOCA* condition to the level of the mode strategy VOCA+Gest(r) (the only redundant mode strategy that is unaided). (The improvement is still not enough to attain or surpass the level of the other redundant mode strategies). Also, with bottleneck reduction, the ranks of joint aided-unaided mode strategies VOCA+Voc(r) and VOCA+Gest(r) are the same, whereas without it, VOCA+Voc(r) is the stand-alone second-ranked strategy (see figure 7.7).

Figure 7.8 Mode strategy profiles for the *UI VOCA* and *MM2 VOCA* conditions when value function V_1 is used.



Discussion: What does MSIM demonstrate about the bottleneck reduction hypothesis? The central question of the present investigation was formulated in chapter 4: bottleneck reduction will, very likely, improve a communicator's aided-unimodal strategy, but will it improve the communicator's repertoire of mode strategies as a whole? A better understanding is needed of the relationship between, on the one hand, the individual's communicative resources and the interface of the VOCA and, on the other, the repertoire of mode strategies that they afford to that individual. The *MM2 VOCA* condition characterized a VOCA that implements bottleneck reduction (i.e., one that has a multimodal interface and that has an improved mode of synthesized speech).

MSIM demonstrates that, even without bottleneck reduction, the aided-unimodal strategy is certainly not the best mode strategy for a communicator to use in scenarios in which the communication partner is moderately or very familiar;[16] instead, the unaided mode strategies are best (the aided-unimodal strategy is ranked last by all of these values functions and the unaided mode strategies are ranked highest). For these scenarios, MSIM also demonstrates that bottleneck reduction *does not* improve the value of the aided-unimodal strategy. The reason is that the gains that are produced by bottleneck reduction are not valued very highly when the aided-unimodal and joint aided-unaided mode strategies are evaluated.

MSIM demonstrates that, in scenarios in which the communication partner is unfamiliar and the communicator's tolerance to fatigue is high,[17] the mode strategies that entail redundancy are most highly ranked, followed by the aided-unimodal strategy, VOCA(nr), then followed by the non-redundant and unaided mode strategies. However, MSIM demonstrates a way in which the communicator's tolerance to fatigue can play a role in how the mode strategies are evaluated; if the tolerance drops to moderate from high,[18] then the value of the aided-unimodal strategy drops as well. Since the lowered overall value of the aided-unimodal strategy drops below the values of the non-redundant and unaided mode strategies, MSIM shows that a shift in the communicator's tolerance to fatigue can have an impact on the mode strategies that the communicator decides to use.

MSIM demonstrates, for scenarios in which the communication partner is unfamilar and the communicator's tolerance to fatigue is not low, that bottleneck reduction *does* improve the overall value and ranking of the aided-unimodal strategy. However, this gain is tempered by the fact that other mode strategies are more highly ranked anyway. The number of more highly-ranked mode strategies depends on the communicator's tolerance to fatigue. If the tolerance is high, then bottleneck reduction moves the aided-unimodal strategy into the top three (including ties), and if it is moderate, then the aided-unimodal strategy moves into the top four (including ties). The latter is a more drastic improvement than the former, however.

Also, MSIM demonstrates that bottleneck reduction improves the value of most of the joint aided-unaided mode strategies, but has a negative effect on the joint aided-unaided mode strategy VOCA+Voc(r). This effect can be completely attributed to increased mode conflict.

It is difficult to determine whether the negative effect on the mode strategy VOCA+Voc(r) corresponds to a *global* detriment with respect to the entire set of joint unaided-aided redundant mode strategies (or even the entire repertoire of mode strategies) because there are several different ways

[16] In MSIM, in such scenarios the value functions that is used would be one of V_3, V_4, V_4, V_5, or V_6

[17] In such a scenario, MSIM would make use of the value function V_1.

[18] In such a scenario, MSIM would make use of the value function V_2.

to characterize global effect. If we base our characterization on the differences with respect to the rankings of the mode strategies between conditions, then bottleneck reduction does not result in a global detriment. However, if we base our characterization on the differences with respect to the overall values of the mode strategies that belong to an illustrative set (namely, the set that consists of the aided-unimodal strategy and the joint aided-unaided redundant mode strategies that make use of two modes; see above for the rationale for the use of this particular set), then MSIM demonstrates that bottleneck reduction does result in local improvement (i.e., with respect to the aided-unimodal strategy), but global detriment. In either case, MSIM demonstrates that the improvement to the aided-unimodal strategy does have a cost, and that cost is exacted on a more highly-ranked mode strategy.

If improvement to the aided-unimodal strategy does have a cost, and this cost is exacted on a more highly-ranked mode strategy, what does this mean in practical terms for an aided communicator? According to MSIM, if an individual uses a VOCA with a unimodal interface in a scenario with an unfamiliar communication partner, he or she would find the use of synthesized speech augmented with vocalizations to be better than synthesized speech alone. Because the interface is unimodal, the vocalizations can be performed simultaneously (e.g., the vocalizations could modulate the meaning of the synthesized speech, perhaps to add emphasis). MSIM illustrates this to be an advantage over the use of synthesized speech and gesture (which cannot be used simultaneously, since the modes conflict). MSIM demonstates that even if the individual instead were to use a VOCA with a multimodal interface, such as the hypothetical VOCA that was described in section 7.2, it would be the case that both the mode strategy of using synthesized speech and vocalization together and the mode strategy of using synthesized speech alone would be improved, but it would still be better to use the former mode strategy than the latter. Except now, when the former mode strategy is used, the simultaneous use of synthesized speech and vocalization is not possible due to mode conflict, and the communicator would be instead forced to use the modes in sequence. Now, is the loss of the ability to use modes simultaneously truly offset by the improvement to the synthesized speech for aided communicators, which are a hetereogenous group, and for other joint activities? Future research can investigate this line of inquiry, now that this tradeoff has been identified, formalized, and computationally demonstrated.

7.5 Summary

This chapter described the simulation results produced by MSIM: (1) in scenarios in which the communication partner is familiar, the unaided strategies are optimal; (2) the mode strategies that make use of the aided mode of synthesized speech are not optimal unless their relatively high cost in terms of physical effort is outweighed by the need for improved interpretability; and (3) even when the expense of the aided mode of synthesized speech is justified, it is better to use that mode in combination with the other, unaided modes, such as gesture and vocalization. Empirical and anecdotal evidence agrees with these simulation results.

This chapter described a simulation condition for MSIM that implemented *bottleneck reduction* — that is, a VOCA was defined that afforded a mode of synthesized speech that is improved through the incorporation of a multimodal interface (that is, improved over the synthesized speech which

is afforded by a VOCA with a unimodal interface). MSIM was used to demonstrate the impact of bottleneck reduction by means of a sequence of comparisons between pairs of contrasting simulation conditions. In the first comparison, a negative effect was demonstrated on the domain-goal-specific values of one of the joint aided-unaided mode strategies, VOCA+Voc(r), due to increased mode conflict. In the second comparison, a positive impact was demonstrated on the domain-goal-specific values of the aided-unimodal strategy, VOCA(nr), and on all of the joint aided-unaided mode strategies, due to improvements on the mode of synthesized speech. In the final comparison, a positive effect on the aided-unimodal strategy, VOCA(nr), and a negative effect on the set of joint aided-unaided mode strategies were demonstrated, due to the collective impact of increased mode conflict *and* an improved mode of synthesized speech.

Thus, MSIM demonstrated, first, a way in which different VOCA interfaces afford different repertoires of mode strategies, and, second, a way in which bottleneck reduction might afford *local* improvement, but not *global* improvement.

Chapter 8

Conclusion

This chapter summarizes this dissertation and outlines some aspects of this research to be pursued in future work.

8.1 The AAC Design Dilemma

A goal for AAC researchers is to develop clinical interventions for individuals who have little or no functional speech, gesture, or writing. A wide variety of interventions have been developed, many of which are computer-based and provide synthesized speech. Improvements and advances in input devices and interface hardware promise the possibility of new and better interaction styles with these computational devices — specifically, the use of multiple modes of input has great potential for improving Voice Output Communication Aids.

We showed in chapter 2 that there is an important distinction between the modes of communication and the modes of articulation in an individual's repertoire: the mode of synthesized speech is *not* a mode of communication, but rather a mode of articulation. In chapter 3, and in the context of having made this distinction, we described the AAC research literature that showed that the so-called *unaided* modes in an aided communicator's repertoire (such as vocalization, facial expression, gesture, and gaze) are important because they afford unimodal and multimodal strategies. We also showed that the so-called *aided* mode in an aided communicator's repertoire (the mode of synthesized speech that is provided by the VOCA) affords the unimodal aided strategy of synthesized speech. We described the empirical and anecdotal evidence that shows that the mode strategy that an aided communicator uses varies; the type of communication partner and the demands of the communicative scenario are thought to be the most influential factors.

We can see that an aided communicator is faced with multiple, conflicting goals when producing his or her multimodal communicative actions. For instance, it is important that the action be both *effective* (that it accomplish its intended purpose) and *efficient* (that it require the use of as little physical effort as possible). But the highly effective actions are often the ones that require a great amount of physical effort to produce. Clark [1996] formalized the simultaneous persuit of effectiveness and efficiency (as well as social politeness) as the desire to satisfy different *types* of goals. He identified three such goals: domain, procedural, and interpersonal. We hypothesized that commu-

nicative exchanges that are unmediated (e.g., "typical" face-to-face conversations) and those that are AAC-system-mediated are similar in that their interlocutors attempt to satisfy these three types of goals. Moreover, we showed in chapter 3 that the conflicts that can occur among these three goals is exacerbated for aided communicators. By applying and extending Clark's model of communication to AAC-system-mediated communication, we hypothesize that the relative importance of these goals varies, and aided communicators choose the multimodal communicative action that best satisfies these goals. The variation in the use of mode strategies by a communicator in various situations demonstrates that the mode strategy that has the best combination of effectiveness and efficiency at any given point in time changes. In some situations, the unimodal or multimodal unaided mode strategies are best; in others, multimodal strategies that combine the aided and unaided modes are best, and in yet others, the unimodal aided mode strategy is best. It can also be seen that in some scenarios, the unimodal aided mode strategy is the *only* one that can be effective, and we hypothesize that this trumps any consideration for efficiency.

The application of Clark's theory also suggests that we need to recognize that aided communicators have different types of goals. More specifically, they have domain, procedural, and interpersonal goals. We applied this model in the explanatory mechanisms that were developed in this work.

A multi-pronged intervention strategy for communication disorders is therefore warranted — provide therapy to improve an individual's unaided modes to the greatest degree possible, and develop computational VOCAs to improve the individual's aided mode. An individual's overall repertoire of mode strategies will be thereby improved. In chapter 4, we identified a previously-unacknowledged interaction between the aided and unaided modes. The starting point was the observation that the use of the VOCA requires input actions that are produced using the same communicative effectors that support the use of the unaided modes. The simultaneous or even sequential use of some modes might be precluded if the support required by each exceeds what can be provided by the underlying communicative effectors. We described this situation as *mode conflict*. The repertoire of mode strategies that a repertoire of modes affords is reduced by conflict among the modes.

More than a decade ago, Shein et al. [1990] proposed what we have termed the *bottleneck reduction hypothesis*, which holds that increasing the information throughput of the interface of VOCAs (or, stated conversely, reducing the bottleneck that presently exists there) will bring about a more effective AAC system, and that this would be best accomplished through a VOCA interface that is multimodal, rather than the unimodal ones that are presently used. In essence, a VOCA with a multimodal interface would afford a better aided mode of synthesized speech than one with a unimodal interface. And researchers such as Roy et al. [1994a], Keates and Robinson [1998], and others have developed preliminary versions of such multimodal interfaces. But Shein et al.'s analysis of the consequences of bottleneck reduction did not include in its scope any potential impact on mode strategies other than the unimodal aided one. In chapter 4, we showed that any potential improvement to the unimodal aided strategy, achieved through the development of a multimodal VOCA interface, *would* have a detrimental effect on at least some of the multimodal strategies: the input actions required to use a unimodal interface conflict with the support of the unaided modes, and the input actions for a multimodal interface would conflict even more. Thus, a multimodal interface would have a detrimental effect on an interlocutor's use of strategies that combined the

use of aided and unaided modes. Is this tradeoff justified? It might be, provided that the unimodal aided strategy, thus improved, becomes the best one to use, even in situations in which the negatively-affected multimodal strategies might otherwise have been best. It might not be justified, however, if there remain situations in which the aided communicator still requires the now negatively-affected multimodal strategies. The issue of mode conflict is clearly relevant to AAC design, but has not been previously acknowleged, let alone investigated.

We decided, therefore, that a formal account was required of the mechanisms that are relevant to the intertwined processes of interacting with a VOCA and being engaged in a communicative exchange — a process that is partially mediated by the device. We wanted to account explicitly for the variation in mode strategy selection in different situations and for the advantages and disadvantages of unimodal and multimodal VOCA interfaces in those situation in a way that could be demonstrated computationally.

8.2 Contributions of this dissertation

8.2.1 The Characterization of Foundational Notions

In the course of conducting this work, we developed definitions for the foundational notions of *mode*, *modality*, and *communication channel*. These definitions are the most thorough that we know of, and reflect an analysis that took into account many different sources. Although the provision of definitions can be a routine matter in a research project, these denotations of these particular terms have been inconsistent and problematic in the research literature.

In chapter 3, we also examined the notion of a *communication disorder*. We applied the model developed by Clark [1996] that characterizes communication as joint activity, and showed that it is theoretically incongruous to describe an individual as having a communication disorder.

8.2.2 Analysis of AAC Interventions

In chapter 2, we examined three different models of communication, and we showed how each of these models implies its own model of dysfunction in communication. We related each model of dysfunction to the AAC literature. The investigation into the various models of dysfunction was motivated by the realization that they provide the rationales for the strategies for intervention and the designs of AAC systems and VOCAs, and that inaccuracy or misrepresentation in the underlying model has the potential to be realized as an ineffective intervention strategy.

We found that the model of dysfunction that we derived from Clark's theory of communication as joint activity is novel and is not found in the AAC literature. This model generates hypotheses about some potential sources of dysfunction, such as an aided communicator's inability to participate in the establishment of a domain goal, and a communication partner's misinterpretation of an aided communicator's lack of response as an indication of understanding. These sources might be further investigated in future work and might even generate new proposals for alternative approaches to intervention.

From the application of Clark's theory, we developed a conceptualization of AAC systems as serving to mediate joint activities, albeit partially. It follows that aided communicators require an

adequate repertoire of mode strategies to meet the various and changing demands of joint activity and that a single mode strategy, such as the unimodal aided strategy, is not likely to suffice. This conceptualization stands in direct contrast to another one, one that does have currency in the AAC literature — that AAC devices serve as voice prostheses. This is not an adequate view because it implicitly locates the source of the dysfunction in communication in the process of speech production. But the production of speech or any other communicative action takes place in a context — that context is the activity in which the interlocutors are collaboratively engaged. The process of communication *emerges* from the underlying process of joint activity. Locii of dysfunction can be identified in the underlying process as well.

8.2.3 Analysis of AAC Design Process

In our analysis in chapter 3 of the process that is followed in designing AAC systems, we introduced the notions of *local* and *global* effectiveness in order to distinguish between the effectiveness of an AAC system in a particular communicative scenario, as opposed to its effectiveness in a set of different communicative scenarios. This distinction was used in order to establish the connection between the interface of a VOCA and the effectiveness of the AAC system of which the VOCA is a component. We argued that global effectiveness ultimately depends on the utility of the individual's repertoire of mode strategies. That is, an AAC system is effective if it affords at least one mode strategy in the interlocutors' repertoires that is appropriate for the present communicative scenario. Later, in chapter 4, we showed that while modifying the interface of a VOCA from unimodal to multimodal might have a positive effect on the unimodal aided mode strategy, it might also have a negative effect on other, multimodal strategies. An account of which mode strategies are most effective in the various types of exchanges is still an open research question. The true consequences of a multimodal VOCA will be revealed only by an evaluation of its global effectiveness (an evaluation of local effectiveness could be biased toward unimodal aided strategies and fail to reveal the impact of negatively-affected multimodal strategies). It is premature to state with certainty that this tradeoff is a good one or poor one; we feel that this tradeoff merits further consideration.

The AAC design process is iterative, and it is clear that the evaluation of AAC systems plays a crucial role. A lack of feedback about the effectiveness of various systems is a detriment to the iterative design process. And yet, our analysis suggests that yet another type of feedback is needed — feedback with respect to the impact an AAC system has on an individual's repertoire of mode strategies. This only exacerbates the problem. To address this, we proposed the development of specialized computational simulation tools. They could be useful to the intervention team, to provide them with feedback on the effectiveness of an AAC design in advance of its actual implementation and use. The simulation tool that we have developed, MSIM, can be seen as a very early, simplified predecessor of such a future simulation tool. (As will be described further below, MSIM is not a predictive model and has not been evaluated as such. The empirical data that is required for such an evaluation is not yet available at the present time. For further discussion of MSIM as it relates to a predictive model, see section 8.3 below).

8.2.4 Development of Explanatory Mechanisms

In this work, we have proposed several mechanisms to explain the variation in mode strategy selection by individuals whose repertoires include both unaided and aided modes. We characterized multimodal communicative actions as a set of temporally-coordinated mode-specific sub-actions. To represent communicative actions as such, we developed a new matrix-based formalism, which was based on the *timeline-based formalism*. This formalism was developed for the representation of both human-produced and computer-generated multimodal communicative actions. We demonstrated the merit of the formalism in a coding study in which it was applied to empirical data.

We defined formally the notion of a *communicative effector*. The communicative effectors are the body parts that are the means of producing the input actions necessary to use a VOCA and provide the support for the use of the unaided modes. We showed that the sub-actions that compose a multimodal communicative action each require the underlying support of communicative effectors. Thus, we showed that the interlocutor's communicative effectors are the resources for the articulation of *both* mediated and unmediated sub-actions. We defined the notion of an interlocutor-specific *support function* in order to characterize these interrelationships formally. We characterized the mode conflict in an interlocutor's repertoire of modes in terms of the characteristics of the support function. The relationship between the support that an interlocutor's communicative effectors provide and the needs of a repertoire of modes and the interface of a VOCA was thus formalized.

Another relationship that we identified and formalized is that between an interlocutor's repertoire of modes and his or her repertoire of mode strategies; this relationship will be further discussed below, in section 8.2.5. The third relationship that we identified is that between an individual's repertoire of mode strategies and the evaluation of the local and global effectiveness of an AAC system. We formalized this relationship in the following way. We first chose a particular joint activity to focus on, the multimodal referential communication task. We developed a characterization of the various states of this task, and then we developed a *value function* to characterize the aided communicator's relative preferences for the various states. Next, we developed an approach for generating a set of candidate multimodal surface realizations; the mechanism takes into account the individual's communicative effectors and the interface of the VOCA. We connected the two by developing a state transition model that predicts, for each of the set of candidates, the state that would result as a consequence of performing the candidate.

We defined equivalence classes among the candidates, each of which corresponds to a particular mode strategy. In order to abstract the characteristics of each mode strategy, we developed a way to characterize such equivalence classes. We described each mode strategy has having a *value* (or *utility*), a measure that syntheses its merit with respect to advancing both the domain and procedural goals (i.e., with respect to both *effectiveness* and *efficiency*). To evaluate the local effectiveness of a VOCA interface, we considered the values of the repertoire of mode strategies thus afforded in a given communicative scenario. To determine global effectiveness, we considered the values of the mode strategies over a set of scenarios.

8.2.5 Analysis of the Bottleneck Reduction Hypothesis

As has already been mentioned, our investigation has revealed a shortcoming in the rationale of the bottleneck reduction hypothesis of Shein et al. One strength of this hypothesis is that it does identify the relationship between the information that the user provides to the VOCA and the quality of the synthesized speech that the user produces through it. But our analysis in chapter 4 shows that the hypothesis is based on the assumption that the communication exchange is wholly mediated by the AAC device, an assumption that does not hold because aided communicators employ a wide variety of mode strategies, both aided and unaided, unimodal and multimodal. We also demonstrated, using the mechanisms described above, that the use of a multimodal interface has several types of consequences: the unimodal aided strategy might be improved, but the repertoire of multimodal strategies that combine the unaided and aided modes is likely to be adversely affected.

8.2.6 Computational Instantiation

We developed computational instantiations of the communicative effectors, the support function, and the repertoire of modes and mode-specific sub-actions in an agent architecture. We also implemented computationally the candidate-generation algorithm and the state-transition model as a decision-making module in the agent architecture. We developed a simulation facility, called MSIM, in which these mechanisms, thus embedded in the simulated communicative agents, could be manipulated. Using MSIM, different parameter values can be specified. These values determine the set of communicative effectors, the support function, and the interface of the VOCA. MSIM also provides a facility that shows the set of candidates and the mode strategies and their values that correspond to the specified parameter values.

But the tool MSIM was not developed as a predictive model of multimodal surface realization and was not evaluated as such. Rather, we used it to demonstrate the explanatory power of the interrelationships and mechanisms that we have developed. The tool demonstrates how modifications to the interface of an individual's VOCA can have an effect, due to the introduction of mode conflict, on the utility of his or her mode repertoire. In addition, MSIM generated a number of hypotheses about the adaptive behaviour of interlocutors in a variety of communicative scenarios.

8.3 Future Directions

The focus of this dissertation was the formulation and computational demonstration of a set of interrelated mechanisms that affect the process whereby an interlocutor derives a temporally-coordinated set of mode-specific sub-actions to realize a given a communicative plan. We showed that these mechanisms explain empirical and anecdotal evidence about the mode strategies that aided communicators emply in various scenarios. These mechanisms are relevant to the design of VOCAs and AAC systems, more generally. Our next step for this work is to incorporate the mechanisms developed in this dissertation into a *predictive model* of multimodal surface realization. An evaluation will be performed on this predictive model that will be more powerful than that which was applied to the set of explanatory mechanisms and that will produce feedback that that will be

used to refine and to improve the explanatory mechanisms. We also plan to relax the simplifying assumptions that were made in this work. And a direction for the longer term is the development of a predictive model of multimodal communicative action (one that incorporates an explicit model of multimodal surface realization) and a predictive model of the outcome of joint activity. Two potential applications of such predictive models will be described below — a computational simulation facility for use in computer-assisted AAC design and an *adaptive* AAC device.

8.3.1 A Predictive Model of the Process of Multimodal Surface Realization

A predictive model of multimodal surface realization generates, given a communication plan and values for the model's various parameters, hypotheses about the characteristics of the multimodal surface realization that the interlocutor is likely to produce. In this thesis work, we have considered the communication plan to be the input to such a model. All other factors and attributes to which the model is sensitive have been considered to be parameters of the model, since they serve to shape the condition in which the model is invoked. We have developed parameters to represent the goals of the underlying joint activity, and the characteristics of the communicative scenario, the interlocutor's own communicative effectors, and the interlocutor's perception of the communication partner. These collectively form what we might call the "invocation condition."

Theoretical Basis

A topic for ongoing research is the validity of the abstraction of the process whereby communicative actions are produced as the three separate processes of communication plan derivation, multimodal surface realization, and motor realization. The research endeavour to derive a predictive model of multimodal surface realization is predicated on this abstraction.

Different interrelationships between these three processes can be hypothesized.

Different models of the production process can be distinguished on the basis of the relationships between the completion of the plan derivation model and the initiation of the realization process, and between the completion of the realization process and the initiation of motor process. For instance, one version is that the process of plan derivation is invoked, followed by the invocation of the process of multimodal surface realization, followed by motor realization. The process of plan derivation implements checking the satisfaction of some criteria that define completion (some of these criteria presumably are conditioned on the formulation of an adequately fleshed-out communication plan), that the satisfaction of the completion criteria triggers the invocation of the multimodal surface realization process, and that the formed communication plan is passed to the multimodal surface realization process as its input. Once a surface realization is derived, it is then passed to the motor processes to be performed. Another model is that the process of multimodal surface realization is always active and derives candidates for the communication plan that is currently active, in whatever form it exists. As the plan derivation process fleshes out the plan, the realization process updates its candidates. When the plan derivation process is complete, whichever surface realization is presently the best candidate gets performed. Yet another model is that the process of multimodal surface realization is always active, and derives candidates for the communication plan that is currently active, in whatever form it exists and motor realization

is triggered even for partially-realized plans. The speaker's percepts of the partially-articulated action and of the communication partner's reaction to it are inputs that feed back into the plan derivation process.

In the work done here, these three processes were abstracted and their interrelationship was abstracted by parameterizing the model of multimodal surface realization. We intended to represent the interrelationship between the two processes through the mechanisms by which the parameter values are set and modified and by which model makes use of the parameter values. It was the case that, in MSIM, these mechanisms were implemented so that the three processes took place sequentially (i.e., the parameter values were assigned, then the process of multimodal surface realization was invoked). In principle, however, the mechanisms for other interrelationships could be implemented with the same parameterization.

Researchers in artificial intelligence and computational linguistics have posited the existence of the process of plan derivation because of its explanatory power. We have argued that positing the existence of the process of multimodal surface realization has explanatory power. But we have also assumed a strictly sequential model. What we need to do next is to relax the assumption of strict sequentiality and show that the explanatory power remains. The mechanisms that we have proposed are highly parameterized; we believe this parameterization will support models other than a sequential one. It characterizes the process of production and hypothesizes the existence of factors independent of plan realization that affect multimodal surface realization. The abstraction also makes it possible to investigate and to model the process of multimodal surface realization independently, to some degree, of theories of communicative action (and models of communicative plan generation), the latter of which being a difficult research problem that has been the focus of research for decades. The explanatory power of this abstraction woul be demonstrated by input-output equivalence.

The explanatory power of this abstraction should be distinguished from its psychological validity. Establishing psychological validity would require a different evaluation, but it is not clear which methodology would suffice. This is left for multi-disciplinary, future work (i.e., collaboration with a cognitive psychologist will be needed). There is motivation for such work, however, as there is presently precious little evidence for or against the psycholinguistic validity of the "mapping" mechanism (that semantic primitives are "mapped to" or "get realized as" observable mode-specific actions — see section 6.4.2).

Evaluation

One of the greatest challenges of developing a predictive model of multimodal surface realization will be its evaluation. We first distinguish between two types of evaluation criteria: input-output equivalence, and psycholinguistic validity. A model that accounts for the types of multimodal surface realizations that are produced by human interlocutors, but not necessarily through the use of the same mechanisms that humans employ, would be described as having *model adequacy* [Chomsky, 1957], *phenomenological validity* [Nass et al., 2000], or, more generally, as having *functional* or *input-output equivalence*. In contrast, a model might account for human behaviours in more than just a superficial way. This has been described as *theoretical adequacy* [Chomsky, 1957], *process validity* [Nass et al., 2000], or, more generally, *psycholinguistic* validity. Clearly, this is a stronger condition

than input-output equivalence.

For the application domains that will be described in the last two sub-sections of this chapter, a predictive model of multimodal surface realization that has input-output equivalence to the human process would suffice. Thus, we focus on this criterion here. Psycholinguistic validity, however, is also of interest because it relates to the broader issue of the psycholinguistic validity of the various models of communication, more generally. The development of an evaluation methodology for this criterion and its application could form the basis for other directions of research.

An extensive review of the literature revealed little previous work with a focus on the evaluation of computationally-generated multimodal communicative actions. For instance, in the *Animated Conversation* simulation [Cassell et al., 1994, p. 416], two animated humanoids (George and Gilbert) performed various multimodal communicative actions in order to complete a short bank-transaction task; however, the purpose of this simulation was to demonstrate that the multimodal actions that the agents produced could be derived from an underlying script that specified the communicative actions symbolically (rather than specifying them in terms of specific actuator positions). The behaviour of the agents was evaluated with an ad-hoc procedure, which readily revealed several problematic characteristics in their simulated behaviour (such as the over-generation of gestures and head nods). Later, in what is one of the more-detailed evaluations described in the research literature, Cassell et al. [2000, pp. 57–60] described what they termed the "lacuna-based" approach to the evaluation of the multimodal communicative actions performed by the embodied communicative agent REA. The approach involved the elicitation of behaviour by REA in a number of different scenarios, which was then evaluated qualitatively with respect to the behaviour by humans that would occur in the analogous situation. Unacceptable deviations from the empirical analogue — described as "lacunae" — were noted; the feedback guided subsequent efforts to develop and refine the architecture of REA. This approach to evaluation, and others like it, were analyzed and described by Baljko [2001b], who showed that this approach confounds the evaluation of multimodal surface realization and the evaluation of communication plan derivation. Moreover, the coding and analysis procedure, the repertoire of scenarios investigated, the characterization of the predicted human behaviours (the empirical analogue), and the characterization of the lacunae, if any, were not reported.

The development of resources and evaluation methodologies for multimodal models and systems is an emerging research area, as evidenced by the nascent LREC workshop series on the topic (LREC2000 and LREC2002) and the working group established by the ACL Special Interest Group on Semantics (SIGSEM) which emerged in 2000. In this direction, we have investigated the possibility of adapting the methodologies that have been employed for the evaluation of *other* types of predictive models (i.e., models of actions other than multimodal, communicative ones). Our investigation revealed the potential utility of the approach of Jones et al. [2000], who evaluated the degree to which the ACT-R cognitive architecture accounted for the effect of cognitive development on problem-solving performance. Our goal is to evaluate the degree to which a predictive model of multimodal surface realization accounts for the effect of different VOCA interfaces and different sets of communicative effectors on the production of multimodal communicative actions. At an abstract level, these goals have parallels, so we plan to adapt Jones et al.'s approach and to apply it in future work.

To evaluate the ACT-R cognitive architecture, Jones et al. identified a task which human subjects *and* a computational agent are able to perform, the Tower of Nottingham block puzzle (a pyramid that is to be constructed out of 21 individual and interconnecting wooden blocks). As a component of their research, Jones et al. conducted an empirical study to gather data about two measures, task completion time and number of block constructions attempted or made, for both child and adult subjects. Thus, *empirical values* for a set of *measures of behaviour* under two conditions (child and adult) were derived, thereby forming the *observed values* against which the model's *predicted values* were evaluated.

In their research, the ACT-R cognitive architecture was embedded in what has been characterized in this work as an agent architecture; the agent architecture also included mechanisms for sensory-perception (visual perception, including simulation of fovea and parafovea) and for motor movement (arm and hand musculature). The agent was then placed in a simulated task environment. (The authors referred to the agent architecture and the simulated environment collectively as the *task simulation* software). One adult version and three different child versions of the agent were derived by specifying different ACT-R parameter values (the different versions each explored a particular ACT-R mechanism, such as the characteristics of working memory, the number of rules available for activation, and the strengths of the various rules; each version was based on a particular theory of development). The adult and child versions can each be seen as a different predictive model.

The authors then performed a set of simulations, which will be referred to here as the *Tower of Nottingham* (ToN) simulations, in which each of the agents performed the ToN task. Thus, predicted values for the two measures of behaviour were thereby derived for each of the different predictive models. For each predictive model, the correlation coefficient and root mean square error were calculated between the predicted and observed values. The authors' analysis showed a good fit between the values predicted by the adult model and the observed values for adults and between one of the three child models. This was interpreted as evidence for the explanatory power of the particular ACT-R mechanism that was manipulated in that particular child model. Thus, the ToN software simulations were used to demonstrate that the ACT-R cognitive architecture successfully accounted for the effect of cognitive development on task performance.

We plan to evaluate the input-output validity of a predictive model of multimodal surface realization by instantiating it computationally, by invoking it under a representative set of conditions, and then by comparing the model's predictions to empirical observations of human interlocutors under the analogous conditions. Future work to develop a predictive model that is based on the basic decision-making module of the agent architecture that was used in MSIM will be described below. But first, we describe an approach for defining the conditions and for deriving appropriate empirical data.

Acquisition of Finely-Grained Empirical Data

The method of evaluation that we have described needs empirical data about multimodal surface realizations that is more finely-grained than that which is presently available. In the few empirical studies that have been conducted, multimodal communicative actions have been categorized by the mode strategy that has been used, but the repertoires of possible mode strategies have been small

and coarsely grained. For instance, the different unimodal strategies have been distinguished (and each coded separately), but all other instances in which multimodal strategies are used are either coded as unimodal or lumped together in one or two categories (e.g., Blischak and Lloyd [1996]; Light et al. [1985b,a]).

In future work, we plan to gather empirical data and to code it at a greater level of detail. We plan to elicit multimodal communicative actions by human interlocutors (the conditions will be described below), and then to code the actions using the timeline-based formalism. With this, the different mode-specific sub-actions and their temporal interrelationships will be represented explicitly. We plan to define measures of behaviour in terms of the characteristics of the mode-specific sub-actions and their temporal interrelationships. It is important that these measures be chosen carefully; values for them must be derivable from the process when it is simulated and when it occurs empirically. One such measure will be the mode strategy employed. We have developed an approach to characterizing the repertoire of mode strategies that is afforded by a repertoire of modes. For a repertoire of three modes (gesture, facial expression, and vocalization), there are at least seven different mode strategies (i.e., the three unimodal strategies, three multimodal strategies of two modes, and one strategy of three modes). Additional strategies can be defined by distinguishing among different patterns of simultaneity and sequentiality.

We have already defined and applied a coding protocol for inventories of gesture, vocalization, and gaze sub-actions. This protocol was based on previous research that characterized inventories, such as Ekman and Friesen's [1978] FACS for facial expression, McNeill's [1992] categorization of gesture, and Kapoor and Picard's [2001] system for gestures of the head (described previously, in section 4.2.2). Investigation into these inventories should be continued in future work.

We plan to investigate different techniques for the evaluation of similarity between two multimodal surface realizations. A very strict approach to defining similarity could be based on token identity (the co-occurrence of the same mode-specific sub-actions, with the same temporal interrelationships to other specific sub-actions). Such a strict approach will probably not be useful, as we don't need the predicted surface realization to be exactly the same as the one produced by a human, just to capture its salient properties. We envision one alternative technique might be based on the minimum edit-distance between the two induced sets of sub-actions. Another technique might be based on the similarity between the temporal interrelationships between the mode-specific sub-actions (these temporal relationships might be represented by a directed graph; so edit-distance between two graphs might be used). A third technique might be based solely on the similarity of the *mode strategy* used in each. A set of possible values for this measure is given by the inventory developed in section 6.4.2. This approach will require developing a measure of similarity between multimodal strategies, and a method for expanding the set of strategies as a function of the domain of possible referents and the number of semantic primitives that distinguish them.

We are presently planning a study to gather information about the multimodal surface realizations that human interlocutors produce under a variety of conditions.[1] We do not consider the communication plan for which the surface realization is being derived to be part of the condition;

[1] As will be described below, the planned predictive model of multimodal surface realization will require a predictive model of the communication partner's interpretation of different multimodal communicative actions. As part of this same study, we also plan to gather data about the communication partner's interpretation of the elicited multimodal referring expressions.

rather, we consider it to be an input to the process and thus it must be controlled. For the reasons described in section 5.3, we believe that the multimodal referential communication task is a good task for eliciting communication plans for referring expressions. In addition, we also intend to investigate the use of sets of potential referents other than the set described in chapter 7. In particular, we plan to include duplicate objects, so that the realization of a spatial semantic primitive, in addition to shape, size, and colour primitives, will be required for its successful identification. In the longer term, we plan to investigate tasks that might be used to elicit communication plans other than those for acts of referring.

To conduct empirical studies, we also must define what is meant by a *condition*. Already we have identified the set of potential referents as one attribute of the condition. In this dissertation, we have also characterized the invocation condition by the goals that are presently active, the familiarity of the communication partner, and the status of the interlocutor's own communicative effectors and their sensitivity to fatigue. We also include the characteristics of the environment in which the exchange takes place and the physical configuration of the interlocutors. The systematic manipulation of these factors is also a direction for future work, as is the indentification of other attributes of the invocation condition.

In this work, we have also implicitly assumed that the process of multimodal surface realization is deterministic. That is, the differences among different realizations can be explained solely in terms of differences with respect to parameter values and inputs. In principle, explanatory parameters exist, even if we have not yet managed to successfully identify all of them. What will the analysis reveal of the data that will be gathered in the empirical study described above? What if the multimodal surface realizations are highly dissimilar, even for the same condition? Does this mean that the process of multimodal surface realization is stochastic, or perhaps it is simply sensitive to inputs or parameters which we have failed to identify and to control for? The data gathered in the empirical study described above will present an excellent opportunity to investigate the validity of the assumption of determinism.

Formulation of State Space

As already described, a direction for future work is to develop a predictive model of multimodal surface realization. We plan to investigate whether the decision-making module in MSIM might serve as a starting place for such a model. The decision-theoretic approach requires that the joint activity be modelled in terms of states and transitions between them; we have developed a basic model of states and state transitions for the multimodal referential communication task. Improving and elaborating both the state space and the state transition model are directions for future work. At present, the states are defined to be the conjunction of three independent attributes: the chooser's degree of fatigue, the listener's mental model of the intended referent, and the chooser's mental model of the intended referent. We related these attributes to the satisfaction of the domain and procedural goals. Although this formulation of the state space has the advantage that the attributes of fatigue and interpretation are modelled separately, we feel that this state space could be improved. First, we assumed that satisfaction with respect to the two goals can be adequately modelled by single attributes. We reasoned that this assumption would restrict the size of the state transition model in order to avoid the computational intractability that is associated with a large

number of states. In future work, more sophisticated models of goal satisfaction should be developed and implemented. Second, we focused on the domain and procedural goals, and did not model the interpersonal goal at all. We reasoned that a model of the first two goals alone would capture many of the tradeoffs and would serve as a tractable starting place. In future work, this simplifying assumption should be relaxed and all three goals should be modelled.

Model of Motor Movement

Recall that the function of the state-transition model is to provide an account of the consequence state that follows a given action. An interlocutor, when deriving a multimodal surface realization for a particular communicative plan, does so in consideration of a number of different outcome attributes. The state-transition model is essential because it identifies the potential consequences of various actions must be identified so that the best possible action can be chosen. In this dissertation work, we have presented initial state-transition models with respect to the outcome attributes of fatigue and intepretation. Improvements and refinements to both of these models are directions for future research.

At present, MSIM makes use of a simple, additive model of physical fatigue; the development of such a model was not a focus of this work, and we did not claim that the model produced high-quality predictions. Rather, the model was developed to demonstrate the overall decision-theoretic mechanism, and we envisioned replacing it with an improved model in the next design iteration. The fatigue model, at present, represents fatigue as an interval value, which is derived from the amount of physical work that is required by a particular action. In future work, we plan to improve this model by representing mode-specific sub-actions in terms of movement features, rather than as atomic units, as they are presently. If movement features were to be used, any particular motor movement could be characterized by a preparation phase and an execution phase. This approach has already been incorporated in the EPIC simulations of Kieras and Meyer [1997] and the cognitive models developed by Ritter et al. [2000] (although their adaptations were not developed for multimodal communicative actions specifically, but rather the human movements relevant to the use of devices such as a keyboard or mouse, or the manipulation of blocks).

Using movement features, the model of motor movement could incorporate empirical findings about the time duration and amount of physical work required for various mode-specific sub-actions. The temporal characteristics of multimodal actions (e.g., time to initiation and duration of physical execution) could be made to reflect a larger set of factors, such as the motor movements previously performed (recently-produced movements are faster than novel ones), the communicative effector that is involved, and the complexity of the movement itself. Such a model could account for the inter- and intra-speaker differences with respect to speeds of hand movements and key-press actions, as well as for the fact that the degree of sensitivity required for a target-seeking motor movement is a function of the size of the target. Empirical findings about vocal motor processes, including both speech and non-speech, could also be incorporated into the motor model. All of these improvements will be needed in order to transform the simple, additive model of physical fatigue that presently exists in MSIM into a predictive model.

In addition, a more-sophisticated model of movement control could be the basis for more-refined models of mode conflict. Presently, we consider only the execution phase of motor move-

ment, although conflicts might also possibly arise during the preparation phase. An account of all of the sources of conflict is especially important for the generation of only viable candidate multimodal surface realizations. MSIM presently generates many candidate surface realizations that are implausible from the perspective of motor movement. With an improved model, we might avoid the generation of some, or even all, of the implausible candidates, thereby improving the validity of the set of candidate surface realizations.

Model of Interpretation

A predictive model of multimodal surface realization also requires a state-transition model that predicts the degree to which a particular multimodal communicative action will be interpreted as intended. The model of interpretability that was used in MSIM was based on the psycholinguistic model of definite reference identification that is based on the notion of a *comparison set* — that adequate information must be conveyed so that the intended referent can be discriminated from the competitors. Clark [1992, p. 103] has argued that this model is problematic for multi-discourse turn referential tasks, because it does not take into account the common ground that may have accumulated. It also has limited application for actions other than those designed for referring. We justified the use of this simplified model by focusing on a particular type of joint activity in which multimodal surface realizations were derived only for such referring actions. Referring expressions are important in many everyday activities, but other joint activities should be modelled too. For future work, the model of interpretation for referring actions should be improved, and subsequently, models of interpretation for other types of communicative actions should be developed.

Presently, the simulated agent that represents the listener does not see the chooser agent and the set of potential referents through simulated vision; its knowledge base is simply updated to reflect information about the potential referents from the start of the simulation and about the communicative actions that have been performed. The model of interpretation would be made more valid if the agent were to sense and perceive its environment through intermediary processes for sensory-perception and memory (as is the case in the EPIC architecture described by Kieras and Meyer [1997] or the ACT-R/PM architecture described by Ritter et al. [2000]).

Multiattribute Functions for Value and Utility

In MSIM, the state-transition model was designed so that it derived a single consequence state for each action. The task of deriving a multimodal surface realization was formulated as decision-making under certainty. Once the state-transition model is improved, as described above, it will derive a *set* of possible consequence states. For instance, the motor movement model may predict a set of possible fatigue effects, and the interpretation model may predict a set of possible interpretations. (This will entail, of course, the derivation of probability distributions over the possible consquence states). Thus, the task of multimodal surface realization needs to be reformulated as decision-making under *uncertainty*. Thus, the multiattribute value function that we developed should be elaborated and developed further into a utility function. For instance, the function should be refined to reflect the interlocutor preferences that are demonstrated in the empirical data that will be gathered. The form of the function, which derives a weighted average of

the two goal-specific value functions using the values w_D and w_P, may require modification.

8.3.2 A Predictive Model of the Process of Multimodal Utterance Design

Given a communication context (which includes the prior discourse history), a predictive model of multimodal utterance design generates hypotheses about the characteristics of the multimodal communicative action that an interlocutor is likely to produce. Such a model would provide the basis for a predictive model of the outcomes of various joint activities; the predictive model of multimodal utterance design could be instantiated computationally in an agent architecture, and two such agents could be placed in a simulated environment and made to perform a given joint activity. As noted previously, Clark's model holds that the process of communication *emerges* from the process of engaging in joint activity. The agent architecture that is defined in MSIM presently has a trivial mechanism for deriving communication plans — it is based on the current state of the underlying joint activity. Although simple, this mechanism is at least consistent with Clark's model.

A model that predicts the outcomes of one or more types of joint activities (or even just certain attributes of them) by participants whose communicative exchange is mediated by an AAC system would be very useful to AAC system design. These attributes ideally would include the degree to which the joint activity was accomplished successfully, the impact, if any, on the perceptions and attitudes of the interlocutors, and the physical fatigue of the aided communicator. The application of such a model will be described in more detail below.

We consider such a model to provide an account of the communication plan that the individual would derive *and* of the multimodal surface realization that would be derived for it. Thus, the predictive model of multimodal surface realization that was described in section 8.3.1 could be incorporated into a predictive model of multimodal utterance design, thereby enhancing its usefulness.

An extensive review of the literature failed to reveal the existence of any previous work that relates to the computational simulation of human interlocutors whose communicative exchanges are mediated by an AAC system (let alone any other type of mediating system, such as computer or video). In fact, only a small number of software simulations have been developed of human interlocutors engaged any sort of communicative exchange or joint task. The *Animated Conversation* [Cassell et al., 1994] and *Tower of Babel* [McIntyre, 1998] simulations are two such examples. The *Animated Conversation* simulations were described above, in section 8.3.1. Their exchange was scripted in advance. In the *Tower of Babel* simulations, simulated interlocutors performed naming, imitation, and discrimination tasks (which were described as games) [McIntyre, 1998]. In order to perform these tasks, they performed "spoken" utterances for each other (and, in order to do so, the agents made use of a shared but evolving language). The purpose of the simulation tool was to illustrate diachronic language change. Other types of computational simulations of humans performing tasks have been developed, but they have not concerned humans acting as interlocutors.

A direction for future work would be to investigate the existing predictive models for plan derivation and to make use of one in combination with the predictive model of surface realization. The integration of these two processes, especially in a non-trivial, non-sequential way, will be an-

other challenge for future work. For instance, when deriving a particular communicative plan, it might be useful to consider the types of multimodal surface realizations that would be available for it. The model of communicative action should make the distinction between different types of goals, which has been a central explanatory mechanism in this work. A model of plan derivation that is based on decision making has been developed by Paek and Horvitz [2000]; in addition, the model developed by Traum [1999] is also based on Clark's model of communication as joint activity. We expect that whichever model of communicative action is to be incorporated, it should be possible to modify it to include a model of the strategies specific to AAC.

One of the greatest challenges for a predictive model of multimodal utterance design will be its evaluation. We discussed, in the previous section, the issue of stochasticity in the process of multimodal surface realization. This issue also is relevant for multimodal utterance design. Intuition suggests that human interlocutors, under the "same" conditions, will not all produce the same particular communicative action. Again, a direction for future work is to explain the variation. Does it reflect the variation and diversity in the relevant sub-processes (such as sensory-perceptual, cognitive, affective, and motor)? Do individual differences explain the variation, and once they are considered, we can see the human interlocutor as deriving his or her action in a way that can be explained in terms of inputs and parameters that are non-random? Are the differences attributable to differences in the communication plans that were derived? Perhaps the differences are attributable to the multimodal surface realizations that were derived for a given communication plan. Is it that the conditions that we imagined to be the "same" were not so, and that as-of-yet unidentified factors were actually of consequence. If these factors cannot be identified, in theory or in practice, then for all intents and purposes, the process should be considered stochastic. Additional empirical evidence is needed in order to address these foundational questions. But if the process of multimodal utterance design in humans is considered to be stochastic, then the notion of input-output equivalence needs to be modified. The output of the model should be considered to be a probability distribution over a set of possible multimodal utterances, rather than one particular utterance. Then the evaluation of validity of the model should be based on the similarity of its predicted probability distributions and empirically-observed ones. Seeing as such empirical data does not presently exist, this is yet another direction for future work.

8.3.3 Computer-Assisted AAC Design

In chapter 3, we proposed the use of computational simulation tools to assist intervention teams with the design of AAC systems. The primary purpose of these computational simulations would be to provide visualizations of the ways in which a particular AAC system will be used, in various scenarios and with various types of communication partners. This would be a useful aid to the intervention team; it could predict at least certain attributes of the outcomes of AAC interventions without them having to be actually implemented and observed in action. These visualizations and predictions could provide the basis for comparing different AAC system designs. We envision that the tool would be especially useful if it could be used to model the design trade-offs with respect to the ways in which the interlocutors (and especially the person with the communication disorder) make use of their communicative effectors.

An advantage of a simulation tool is its low expense. Results for a large number of simulated AAC-system-mediated exchanges could be produced, and for many different configurations of simulation conditions. Although the derivation of these simulation results would still require time and effort by the intervention team (in terms of deriving and specifying the various parameter values), this expenditure would still be less than what is required to develop and to implement an actual AAC system and to evaluate it empirically. Using a simulation tool, feedback might be obtained for *all* of the types of communicative exchanges in the target set. And, in contrast to empirical observation, a simulation tool could provide feedback on many variations of AAC systems, even including novel, unimplemented systems. Moreover, the feedback provided for each system could include information about mode strategy recruitment and degree of mode conflict. Through the use of computational simulations, more designs could be considered by the intervention team and more design iterations could be performed. More-effective AAC interventions might even be developed through this more-intensive design process. Computational simulation tools, however, will require a formal, computationally implementable model of the process of multimodal utterance design.

The intervention team, understandably, might be skeptical about the validity of the simulation results. The skepticism might be assuaged if the tool somehow acknowledged that the quality of its predictions can vary; it could provide a measure of confidence for each of its predictions, thereby differentiating among its predictions. This presents an additional challenge: evaluation will be required to establish the validity of the tool's confidence measures, in addition to establishing the validity of the predictions themselves. We envision that establishing the validity of computational predictions and convincing potential users of this will be a long-term endeavour. It would be a good idea to establish, right from the outset, that simulation-based feedback is not a complete replacement for empirically-derived feedback, when it is available.

8.3.4 Development of Adaptive AAC Devices

In the previous section, we described a predictive model of multimodal utterance design as one of the crucial components of a simulation tool for computer-assisted AAC design. Such a model has another potential use — as a *user model* that might be incorporated into VOCAs themselves. Such a user model would be active throughout the duration of a communicative exchange, generating predictions of which multimodal communicative action is most likely to be performed by the interlocutor at various points (e.g., at the end of each communication partner's conversational turn, the model would generate hypotheses about the aided communicator's *next* communicative action). These predictions would then provide hypotheses as to which modes of articulation would most likely be used; this, in turn, provides information as to which communicative effectors are and are not likely to be recruited. Thus, the user model could provide information about the relative *availability* of communicative effectors during a communicative exchange.

The incorporation of user model in VOCAs is not entirely new. AAC devices have already been developed that have incorporated other types of user models. For instance, the CHAT system made use of a model that predicted the function, in terms of conversational moves, of the user's next contribution; this information was then used by the device to adapt the presentation of the vocabulary

elements to the user [Alm et al., 1992b]. The Floorgrabber system predicted conversational initiative and used that information to adapt whether conversational interjections were made easily accessible or not [Alm et al., 1992a]. The model being proposed here would complement these previous types of predictive models. These previous models have focused on predicting the *type* of communicative action the user might produce, in terms of the function of the action; the user model proposed here would also predict the way in which the user might recruit his or her mode repertoire to realize the functionally-specified action.

A VOCA that is able to exploit information about the relative availability of communicative effectors would have an interface that attunes itself to receive input actions that are performed using whichever communicative effectors are most available — making its interface *adaptive*.[2] Several variants of adaptive interfaces are possible: *unimodal adaptive* interfaces, which are attuned to unimodal input actions, but the mode used might vary; *multimodal adaptive* interfaces, which are attuned to multimodal input action, but the composition of the set of modes might change; and *unimodal-multimodal adaptive* interfaces, which are attuned to both unimodal and multimodal input actions.

We envision that such a prototype adaptive VOCA could initially be implemented using fairly basic components: a modified version of an existing text-composition facility and text-to-speech module, on a tablet computer with a touch-screen, binary input switches, and an additional mode of eyes-free gestural input. The prototype would also need to include mechanisms to gather information about the communicative exchange that is in progress. The device needs to track which multimodal communicative actions the user actually does perform. This way, the validity of its own predictions might be evaluated and adjustments can be made. The device also needs to infer information about the status of the user's communicative effectors (such as level of fatigue) and information about the user's relative priorities of the various types of goals.

In this planned work, it will be important to weigh the benefits of an adaptive device against its costs. Its costs would be manifested in terms of difficulty to be learned and mastered, cognitive load, and the negative consequences that would occur as a result of prediction errors made by the user model. Keates and Robinson [1998] have demonstrated in their pilot study that these consequences are a reality and need to be carefully considered. However, even fairly steep costs might be justified if the benefits are high enough. The intuition behind the design of an adaptive device suggests a solution to the AAC design dilemma that has been described in this dissertation. The device would recruit the communicative effectors to the greatest degree possible, but only to the point that they are not otherwise being used in the support of the unaided modes. An adaptive approach could allow the VOCA to be *selectively greedy* with respect to the consumption of communicative resources, as opposed to the otherwise *straightforwardly greedy* approach that is employed presently.

The evaluation of adaptive VOCAs poses challenges beyond those associated with evaluation in general (these challenges were described in chapter 3). In particular, a contrastive evaluation with a unimodal, non-adaptive VOCA will be needed. The effectiveness of an adaptive VOCA would need to be evaluated both with respect to a variety of communicative scenarios and communicative partners. It is in the context of a large set of communicative scenarios that the benefits of adaptivity are expected.

[2]This approach to design was described earlier in [Baljko, 2001a].

Appendix A

Appendix

A.1 Definition of "Communication Disorder"

The American Speech-Language-Hearing Association [ASHA, 1991, pp. 9–10] defines a *communication disorder* thus: "Individuals with communication disorders are those for whom gestural, speech, and/or written communication is temporarily or permanently inadequate to meet *all* of their communication needs" (emphasis added). A more formal statement of this definition is as follows:

> x has a communication disorder if x is a person such that $\forall$ his or her communication needs, none of a, b, or c is satisfactory (where a, b, and c correspond to gesture, speech, and writing, respectively).

But this definition, strictly speaking, would not apply to an individual for whom *some* — and not *all* — of his or her communication needs cannot be met by the means of gesture, speech, or written communication; thus, such an individual would not be considered to have communication disorder, which was probably not the authors intention.[1]

In section 3.2 on page 26, a modified version of the definition was given: "Individuals with communication disorders are those for whom gestural, speech, and/or written communication is ... inadequate to meet their communication needs." The removal of the words "all of" gives us the desired definition:

> x has a communication disorder if x is a person such that $\exists$ a communication need such that none of a, b, or c is satisfactory (where a, b, and c correspond to gesture, speech, and writing, respectively).

A.2 Characterization of Social Validity

In section 3.3.4 (page 30), Light [1999, p. 14] was cited as stating that the goal of an intervention must have social validity — "it must be considered by the participants and other relevant stake-

[1] For instance, according to this definition, if $\exists$ only one communication need such that one of a, b, or c is adequate, then x would not be considered to have a communication disorder.

holders to be an important behaviour that will legitimately serve to enhance the communication of the participants."

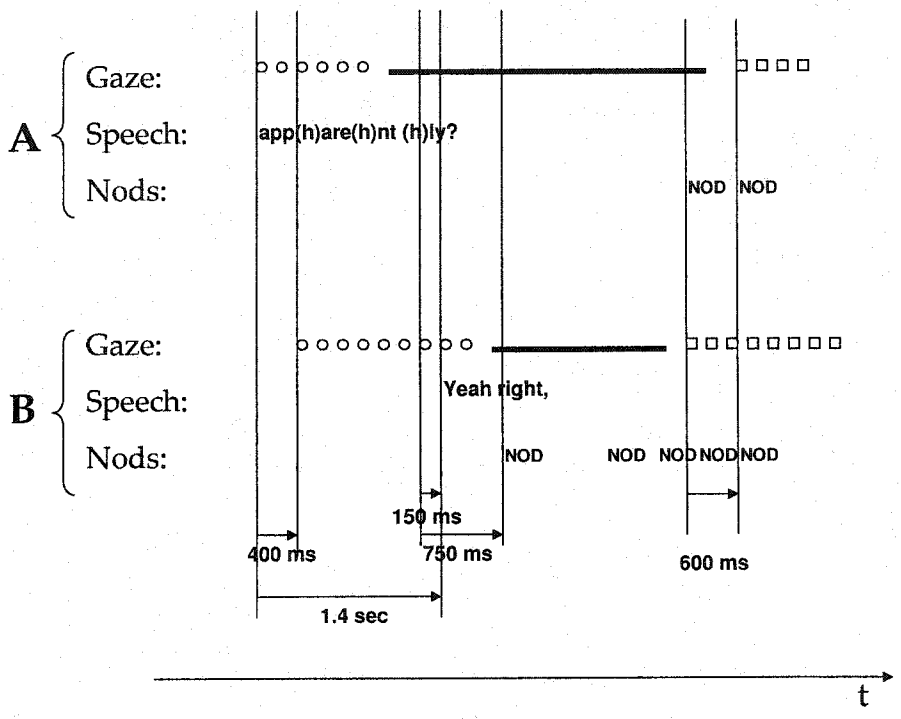
The statement was actually worded as follows [p. 14]: "The intervention goal (i.e., the dependent variable) must be socially valid. In other words, it must be considered by the participants and other relevent stakeholders to be an important and legitimate target behavior that will serve to enhance the communication of the participants."

The definition was reworded to remove the assertion that the stakeholders need to consider the target behaviour to be legitimate — the notion of legitimacy should apply to the enhancement, not the behaviour itself (e.g., "legitimate" and "illegitimate" behaviours are not defined, nor distinguished, by Light [1999]).

A.3 The Timeline-Based Representation Formalism

The matrix-based representation formalism described in section 6.2.4 is quite similar to the *timeline-based* representation formalism that was developed by Goodwin [1981] for his study of co-verbal gaze. An example is given in figure A.1, which shows a portion of an exchange from this study (the illustration was modified for readability by Thórisson [1996, p. 21]). Participants A and B make use of three gaze sub-actions (gaze moving toward other, gaze moving away from other, and gaze at other), two speech sub-actions, and a nodding sub-action (gesture of the head). In Goodwin's representation formalism, mode-specific sub-actions are represented by intervals. The kind of sub-action is represented by the characteristics of the interval — e.g., a solid line, or a line of symbols, such as dots or squares. The exception is speech, which is represented by a text gloss. This is similar to the matrix-based formalism, where the type of sub-action is instead given by the an label that identifies the sub-action from a pre-defined inventory. In Goodwin's representation formalism, the length of the line corresponds to the duration of the action, whereas, in the matrix-based formalism, the number of columns, multiplied by the timestep granularity, represents the duration. In Goodwin's representation, the temporal interrelationships among the sub-actions are explicitly labelled, whereas in the matrix-based formalism, they are implicitly given by the relative column positions of the start and stop times.

Figure A.1 A transcription of a portion of an exchange between two interlocutors, A and B, first prepared by Goodwin [1981, p. 119] and subsequently modified for readability by Thórisson [1996, p. 21]. Dots indicate gaze moving toward the other, a solid bar indicates gaze at the other, and squares indicate gaze moving away from the other. A question mark within transcribed speech indicates rising intonation and "(h)" indicates within-speech plosives.



A.4 Sample Input File to MSIM

```
% Simulation Parameter File - multimodal 2 condition
===< INITIAL INFO >=============================================================================
Agent-L-000
3
0 GazeX 1.00 1.00
1 GestureX 1.00 1.00
2 SpeechX 1.00 1.00
0.00 1.00
Agent-C-000
1.20 1.40 \\ This are the rho_1, rho_2 values
1
1
3
0 Gesture 0.50 0.50
1 Vocalization 0.50 0.50
2 VOCA 1.00 0.10
2 \\ Number of conflicting modes  (FOR modes (prev line), first num is cost, next of IoD
0 2 \\ Gesture and VOCA conflict
```

```
1 2 \\ Vocal and VOCA conflict
0.00 1.00  \\ The agent's tolerance to fatigue and cp fam, resp'y
1
e1
3  \\ number of Semantic Primitives that distinguish e1 (effort, uncert fn, delta min, delta max)
ZSize+(3 1.00 1.00 1.00 )+(3 1.00 1.00 1.00 )+(3 1 1 1 )+(3 1 1 1 )
Colour+(3 1.00 1.00 1.00 )+(3 1.00 1.00 1.00 )+(3 1 1 1 )+(3 1 1 1 )
Shape+(3 1.00 1.00 1.00 )+(3 1.00 1.00 1.00 )+(3 1 1 1 )+(3 1 1 1 )
3 \\ The number of entries to be made to the partial ordering of the semantic primitives
ZSize Colour \\ The primitive ZSize must precede Colour
Colour Shape \\ The primitive Colour must precede Shape
ZSize Shape \\ The primitive ZSize must precede Shape; need to add this because implementation
           \\ transitive closure is not done automatically
===< BODY OF FILE >=============================================================================
121
*****Condition X0Y0
10
Agent-L-X0Y0
3
0 GazeX 0.20 0.31
1 GestureX 0.7 0.31
2 SpeechX 0.96 0.05
0.00 1.00
Agent-C-X0Y0
3
0 Gesture 0.50 0.50
1 Vocalization 0.50 0.50
2 VOCA 1.00 0.10
2 \\ Number of conflicting modes
0 2 \\ Gesture and VOCA conflict
1 2 \\ Vocal and VOCA conflict
0.5 0.5
*****Condition X0.1Y0
10
Agent-L-X0.1Y0
3
0 GazeX 0.20 0.31
1 GestureX 0.7 0.31
2 SpeechX 0.96 0.05
0.00 1.00
Agent-C-X0.1Y0
3
0 Gesture 0.50 0.50
1 Vocalization 0.50 0.50
2 VOCA 1.00 0.10
2 \\ Number of conflicting modes
0 2 \\ Gesture and VOCA conflict
1 2 \\ Vocal and VOCA conflict
0.45 0.55
*****Condition X0.2Y0
[many other condition definitions have been removed]
```

A.5 Additional Simulation Results from MSIM

This section provides additional simulation results from MSIM that augment the discussions in chapter 7:

- Table A.1 provides additional information about the characteristics of the equivalence classes that were produced under the three conditions.
- Figures A.2, A.3, and A.4 illustrate the profiles of agent *C*'s mode strategies in the *unimodal*

interface VOCA conditions for each of the six value functions.

- Figures A.5, A.6, and A.7 illustrate the rankings of agent C's mode strategies in the *unimodal interface VOCA* conditions using each of the six value functions.
- Figures A.8, A.9, A.10, and A.11 each contrast the profiles of agent C's mode strategies between the *unimodal interface VOCA* and *multimodal interface VOCA* conditions (top graph), and also illustrate the rankings that were obtained by MSIM in each of the two conditions (using the maximum overall value of each mode strategy) (bottom diagram), according to the value functions V_3, V_4, V_5, and V_6, respectively (results from the value functions V_1 and V_2 are discussed in chapter 7).

Table A.1 Characteristics of the equivalence sets of candidate multimodal surface realizations that were generated in the *unimodal interface VOCA* and the *intermediate VOCA* simulation conditions (the same candidates were generated in the *multimodal interface VOCA* condition and their goal-specific evaluations resulted in different values). (For the definitions of $\nu(A_i)$ and $\sigma(A_i)$, see section 6.4.2. The sum of the percentages may be slightly incorrect due to rounding error.)

(a)

Characteristics of Γ (the Candidate Set)

Unimodal VOCA

Category Label $\upsilon(A_i)$	Degree of Redundancy $\sigma(A_i)$: 0	1	2	3	
(Gesture) 001	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)
(Vocal.) 010	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)
(Vocal. & Gesture) 011	6 (0.04%)	36 (0.26%)	54 (0.39%)	27 (0.20%)	123 (0.89%)
(VOCA) 100	1 (0.01%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.01%)
(VOCA & Gesture) 101	6 (0.04%)	24 (0.17%)	24 (0.17%)	8 (0.06%)	62 (0.45%)
(VOCA & Vocal.) 110	6 (0.04%)	48 (0.35%)	96 (0.69%)	64 (0.46%)	214 (1.55%)
(VOCA & Vocal. & Gesture) 111	6 (0.04%)	459 (3.32%)	3795 (27.45%)	9162 (66.28%)	13422 (97.09%)
	27 (0.20%)	567 (4.10%)	3969 (28.71%)	9261 (66.99%)	
			13797 (99.80%)		
			13824 (100.00%)		

(b)

Characteristics of Γ (the Candidate Set)

Multimodal VOCA

Category Label $\upsilon(A_i)$	Degree of Redundancy $\sigma(A_i)$: 0	1	2	3	
(Gesture) 001	1 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.02%)
(Vocal.) 010	1 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.02%)
(Vocal. & Gesture) 011	6 (0.10%)	36 (0.62%)	54 (0.93%)	27 (0.46%)	123 (2.11%)
(VOCA) 100	1 (0.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (0.02%)
(VOCA & Gesture) 101	6 (0.10%)	24 (0.41%)	24 (0.41%)	8 (0.14%)	62 (1.06%)
(VOCA & Vocal.) 110	6 (0.10%)	24 (0.41%)	24 (0.41%)	8 (0.14%)	62 (1.06%)
(VOCA & Vocal. & Gesture) 111	6 (0.10%)	321 (5.50%)	1923 (32.97%)	3332 (57.13%)	5582 (95.71%)
	27 (0.46%)	405 (6.94%)	2025 (34.72%)	3375 (57.87%)	
			5805 (99.54%)		
			5832 (100.00%)		

Figure A.2 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* conditions for the value functions V_1 and V_2.

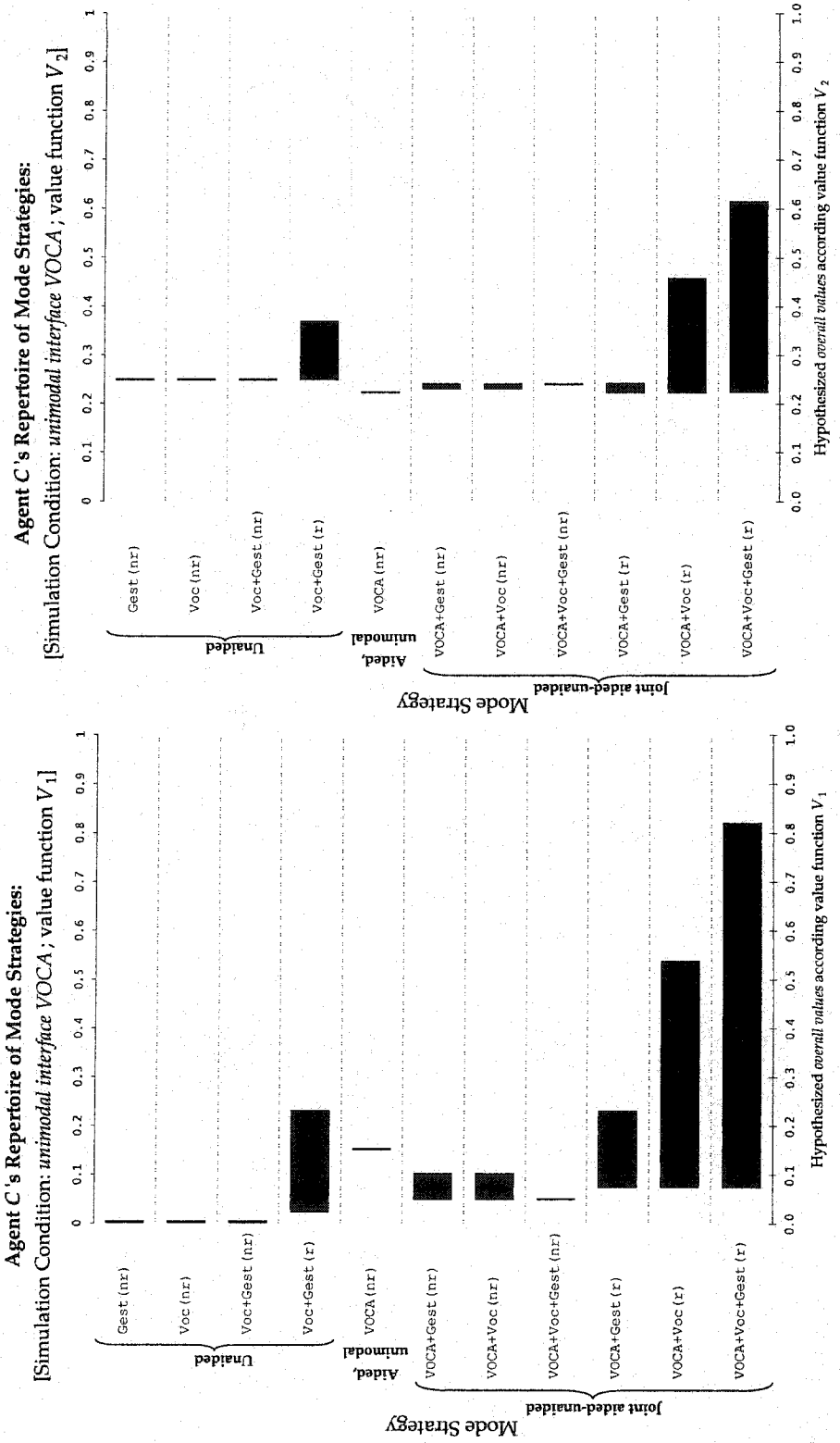


Figure A.3 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* conditions for the value functions V_3 and V_4.

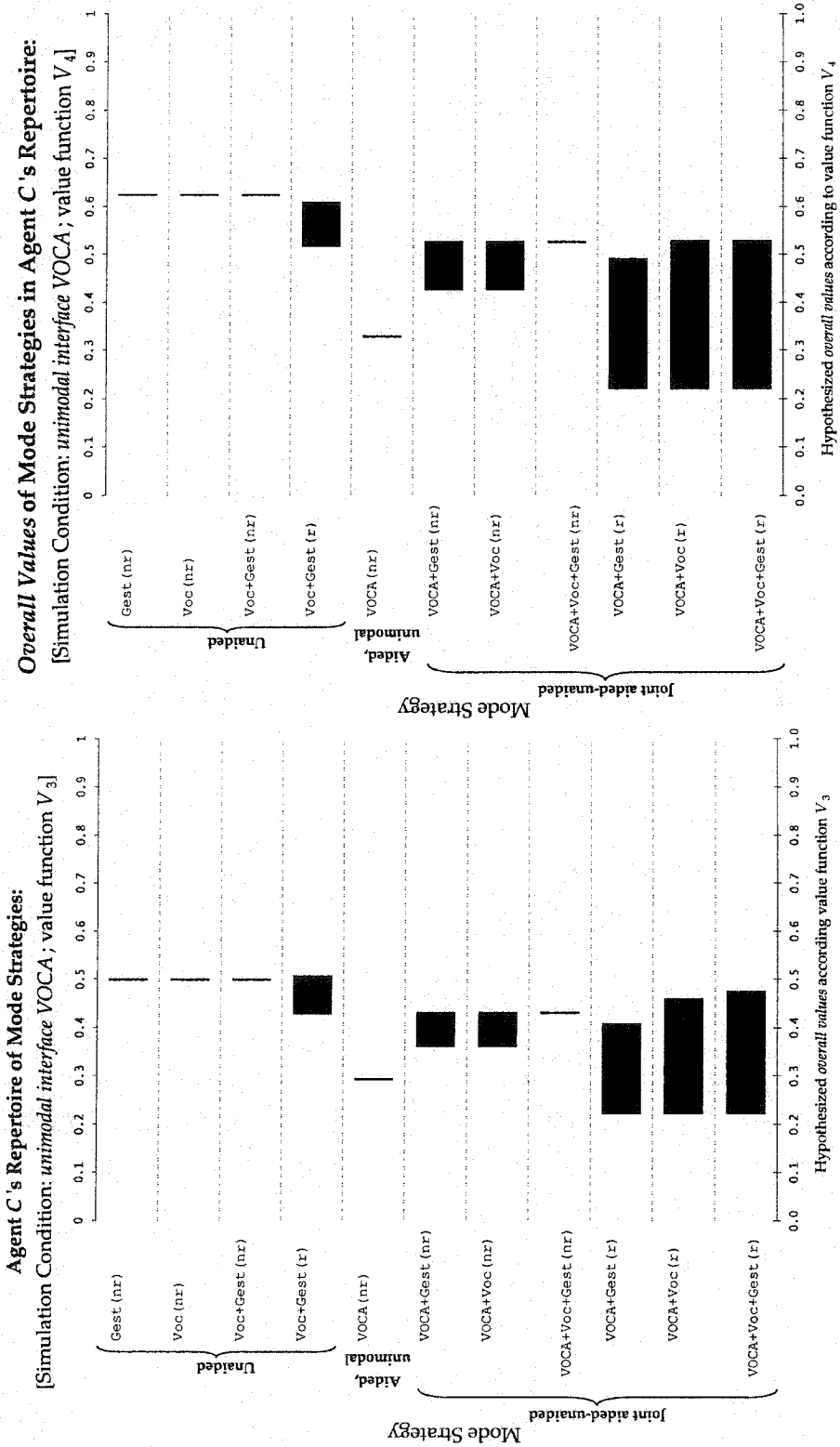


Figure A.4 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* conditions for the value functions V_5 and V_6.

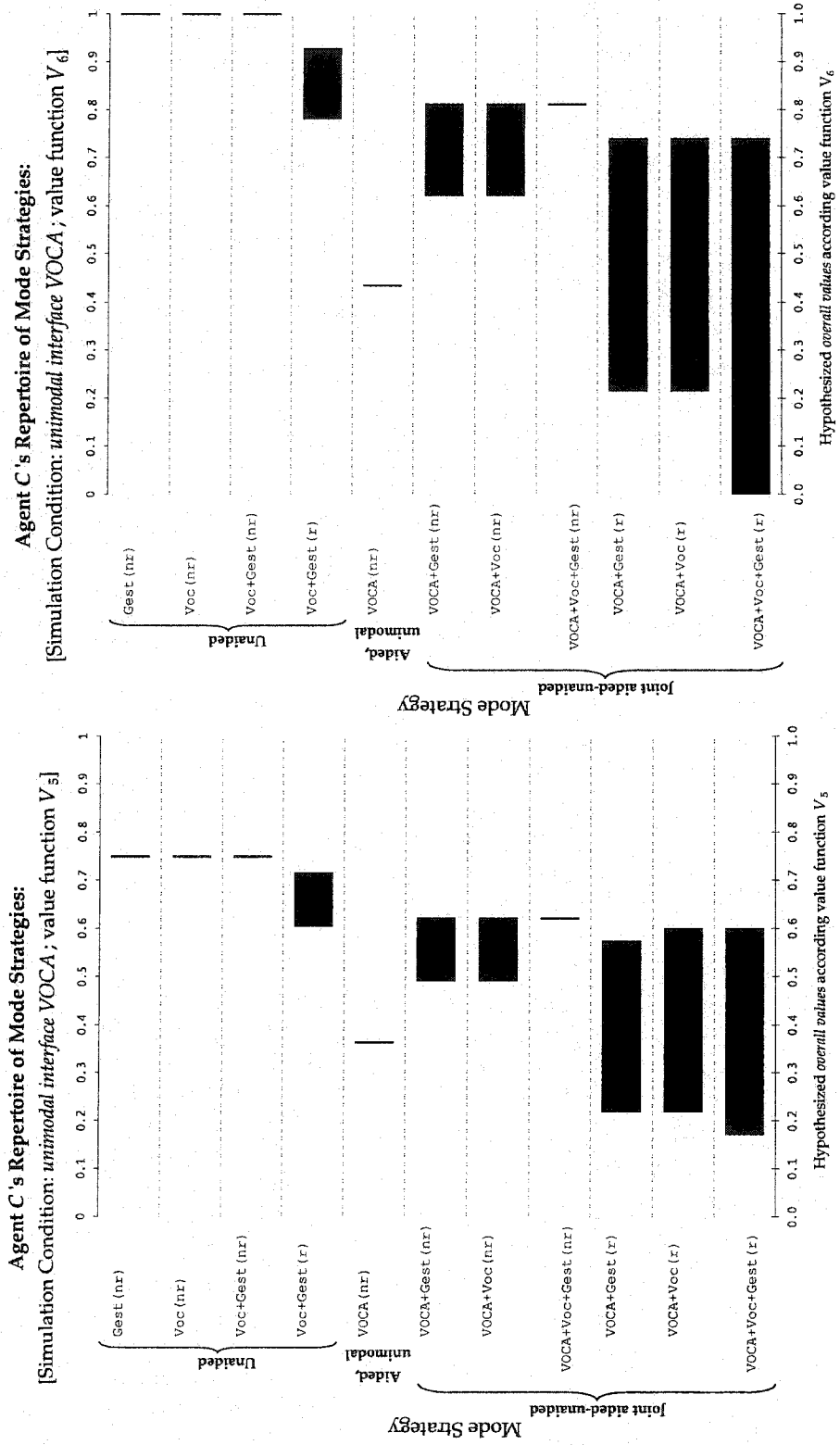
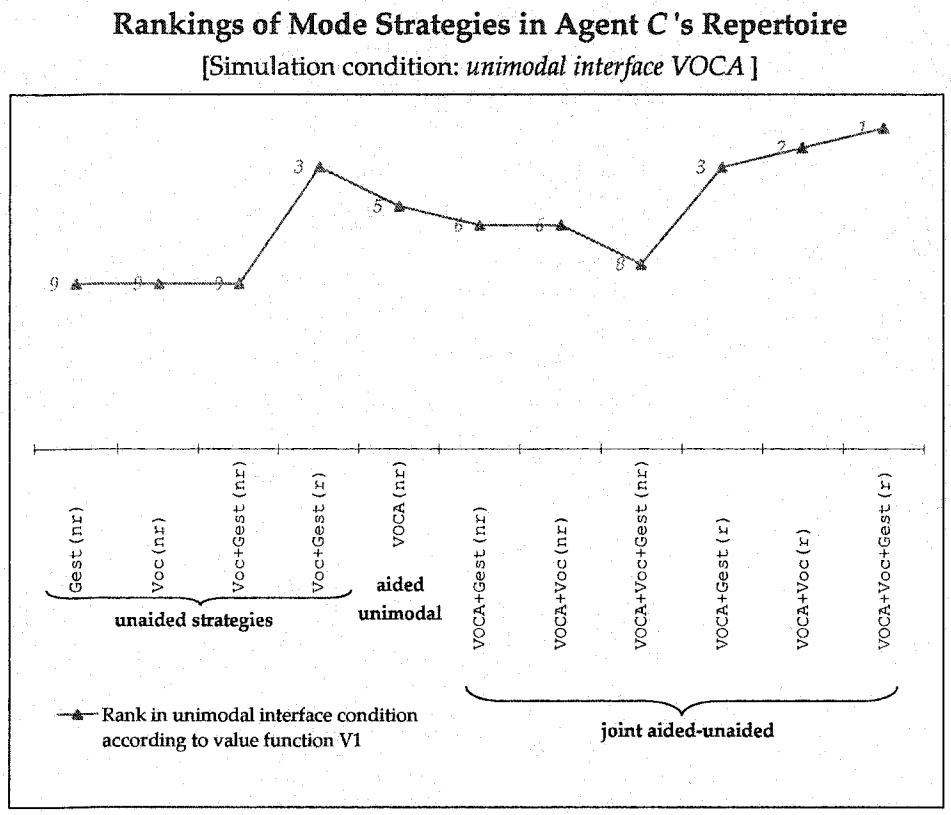


Figure A.5 Profiles of agent C's mode strategies in the *unimodal interface VOCA* conditions for value functions V_1 and V_2.



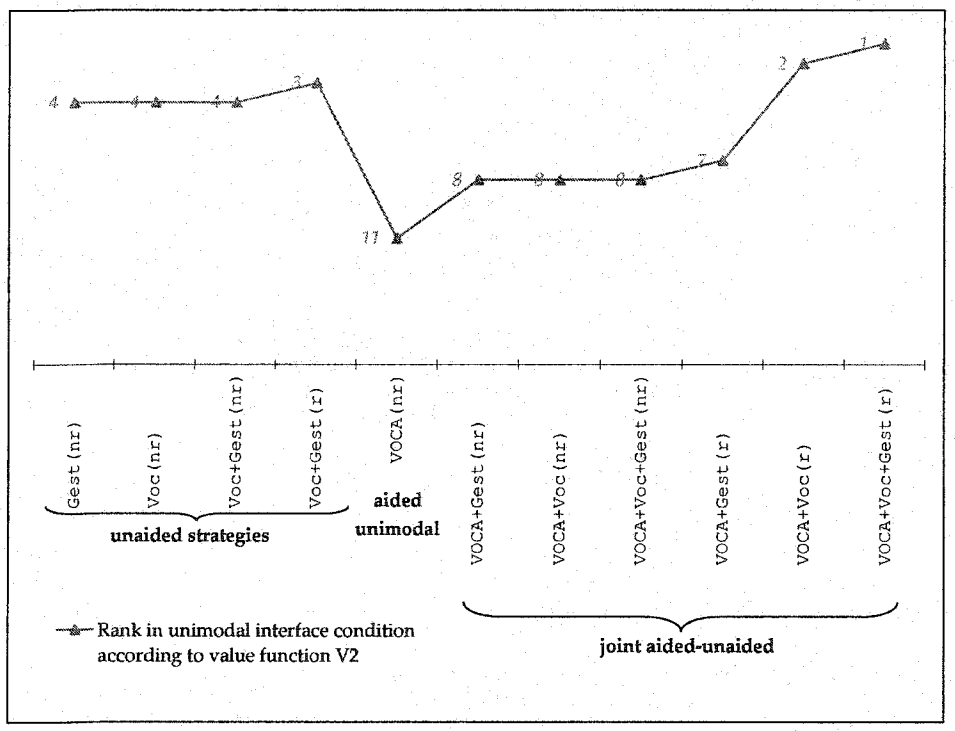
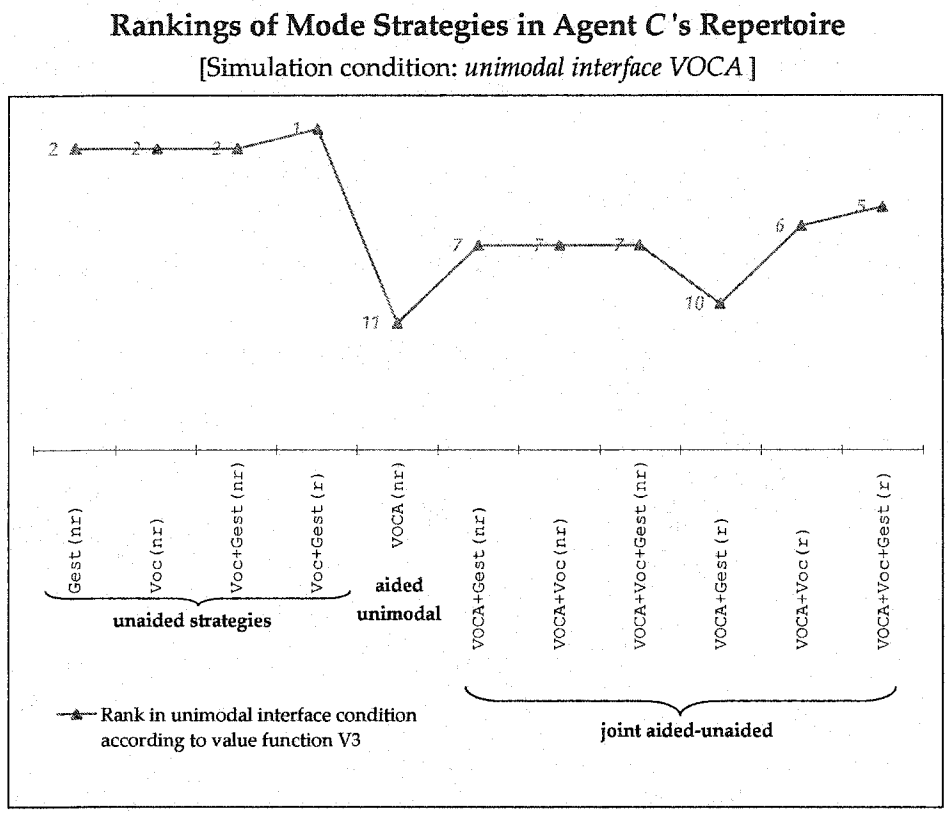


Figure A.6 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* conditions for value functions V_3 and V_4.



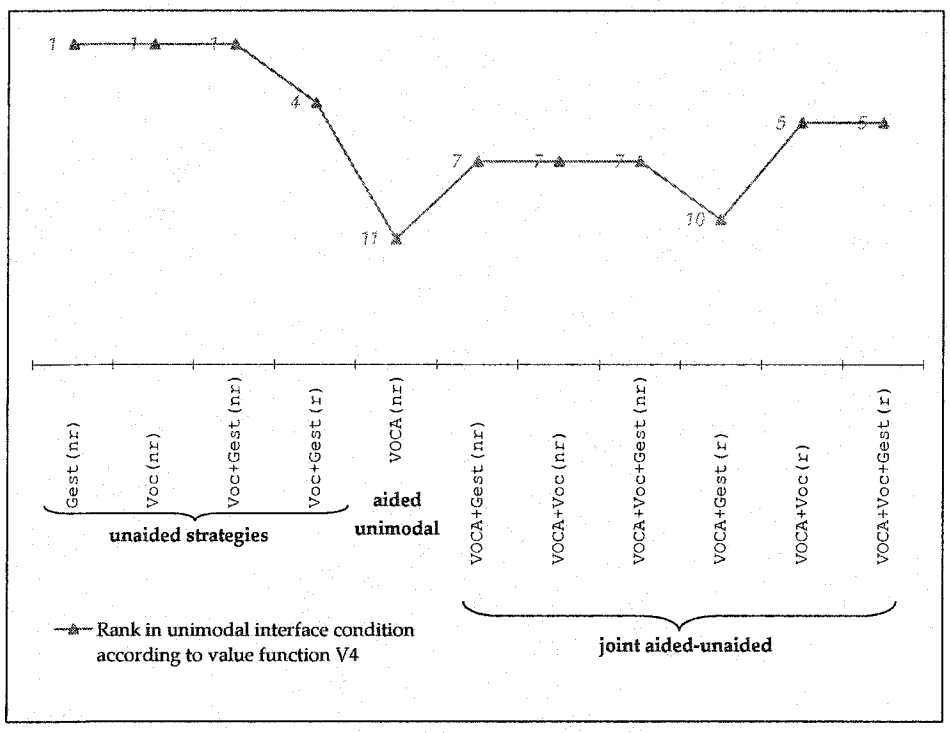
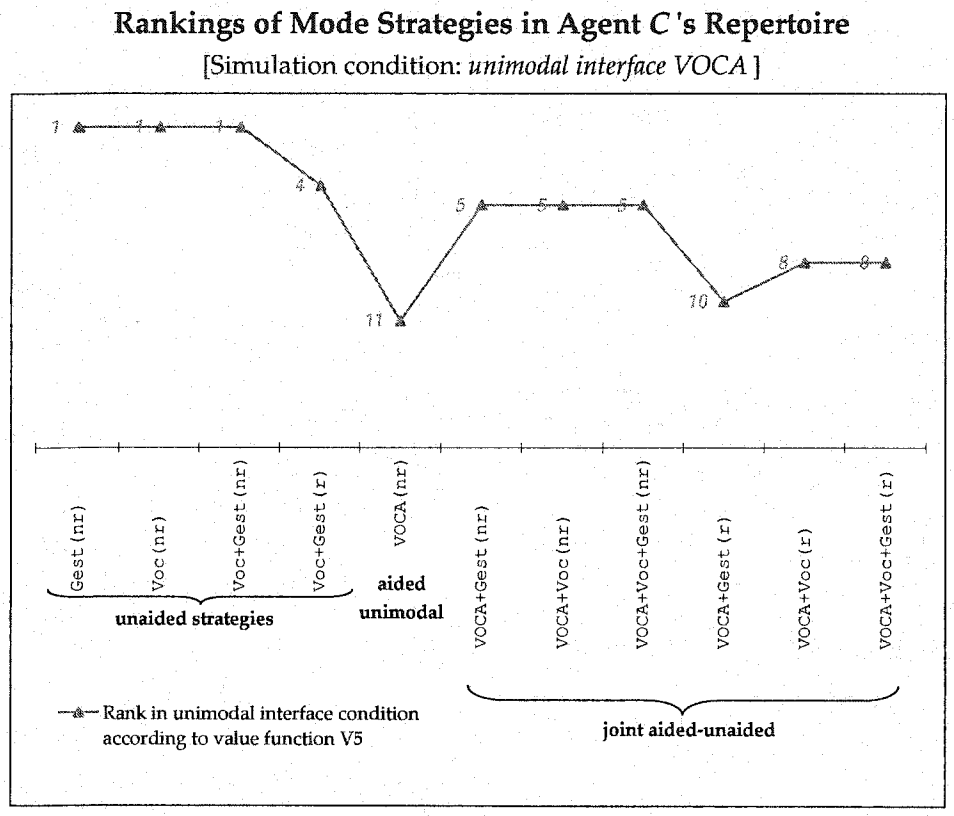


Figure A.7 Profiles of agent C's mode strategies in the *unimodal interface VOCA* conditions for value functions V_5 and V_6.



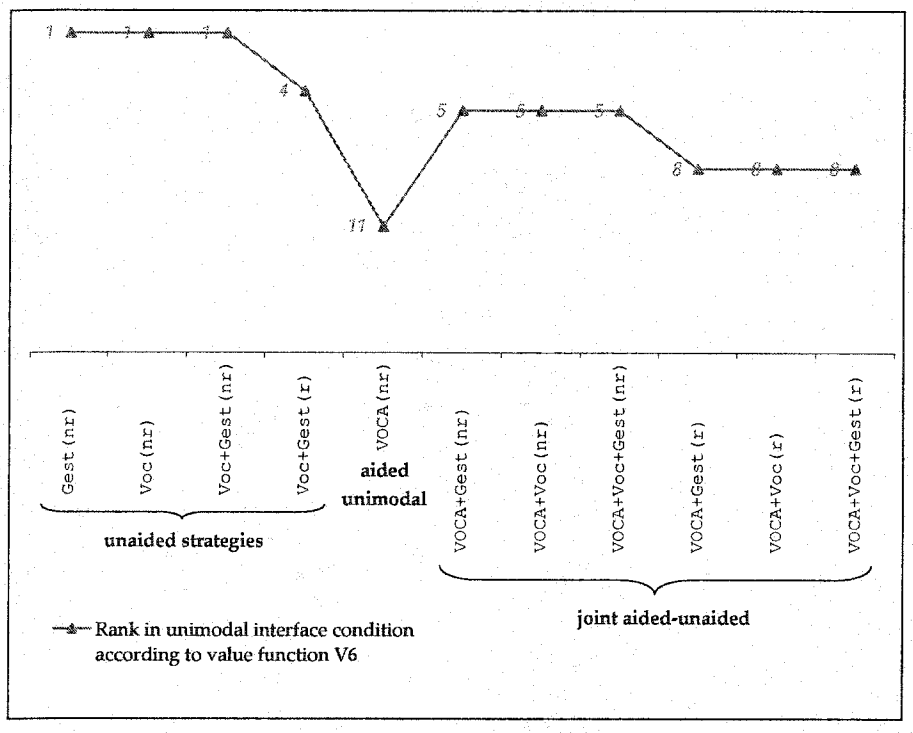
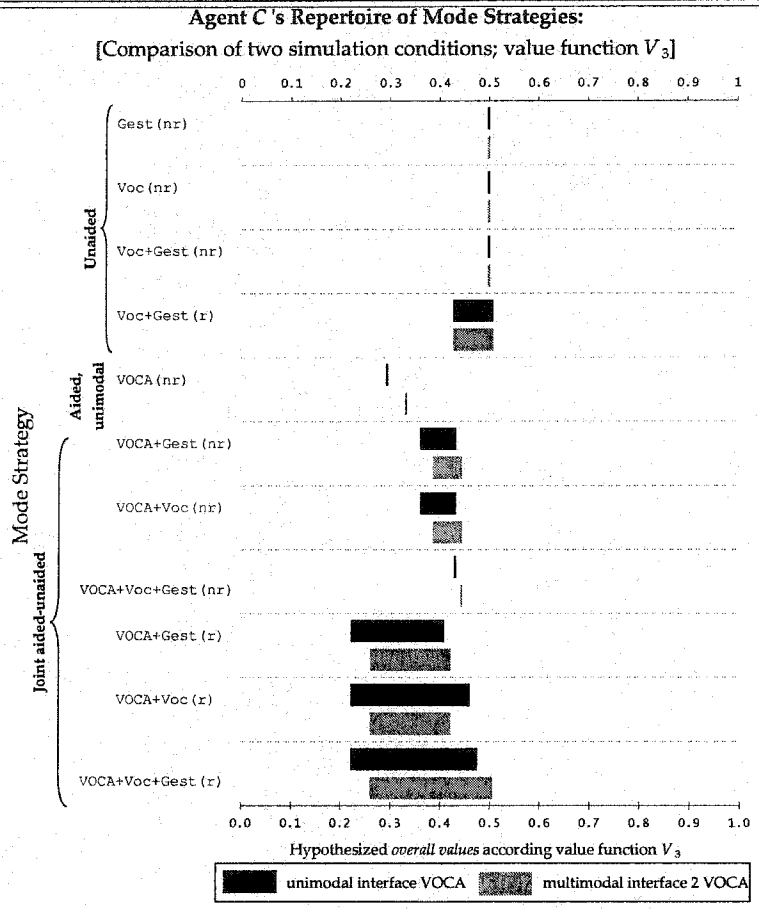


Figure A.8 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* and *multimodal interface VOCA* conditions when the value function V_3 is used (top graph) and the respective rankings that were obtained by MSIM (using the maximum overall value of each mode strategy) (bottom diagram).



Ranking of Agent *C* 's Mode Strategies:

A comparison of the *unimodal VOCA* and *multimodal VOCA* conditions

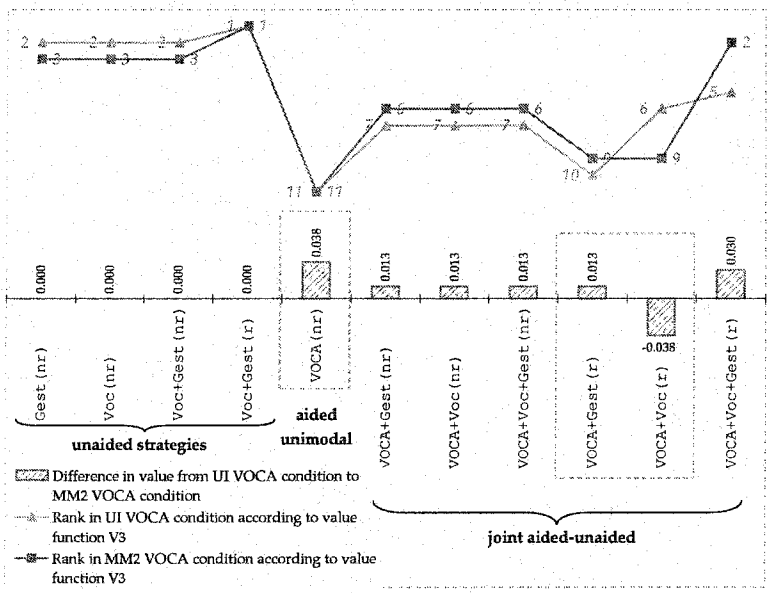
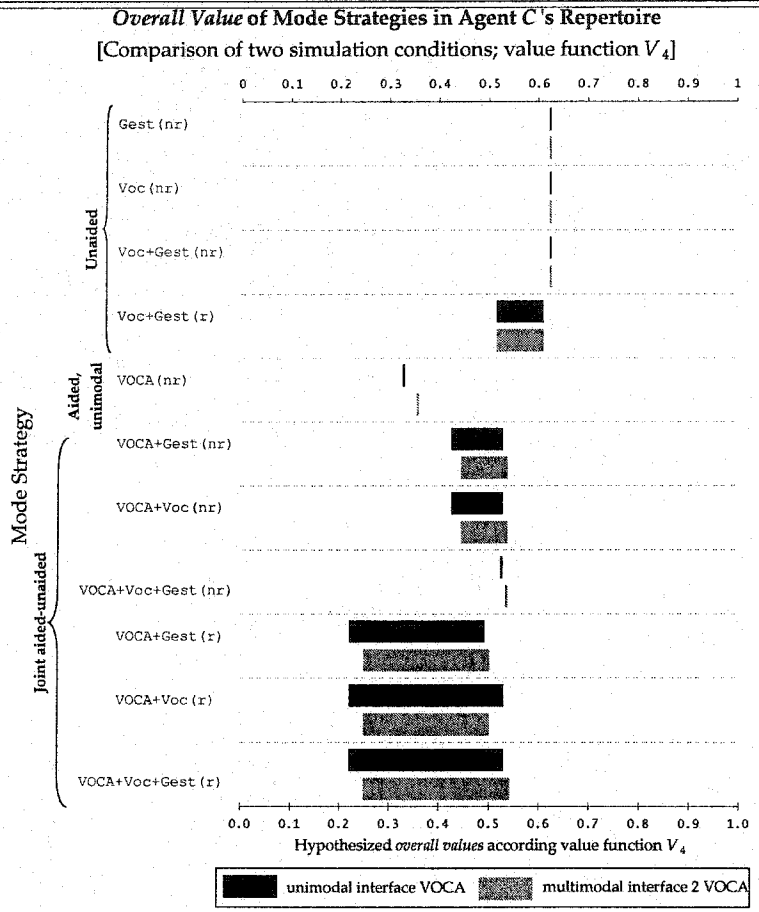


Figure A.9 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* and *multimodal interface VOCA* conditions when the value function V_4 is used (top graph) and the respective rankings that were obtained by MSIM (using the maximum overall value of each mode strategy) (bottom diagram).



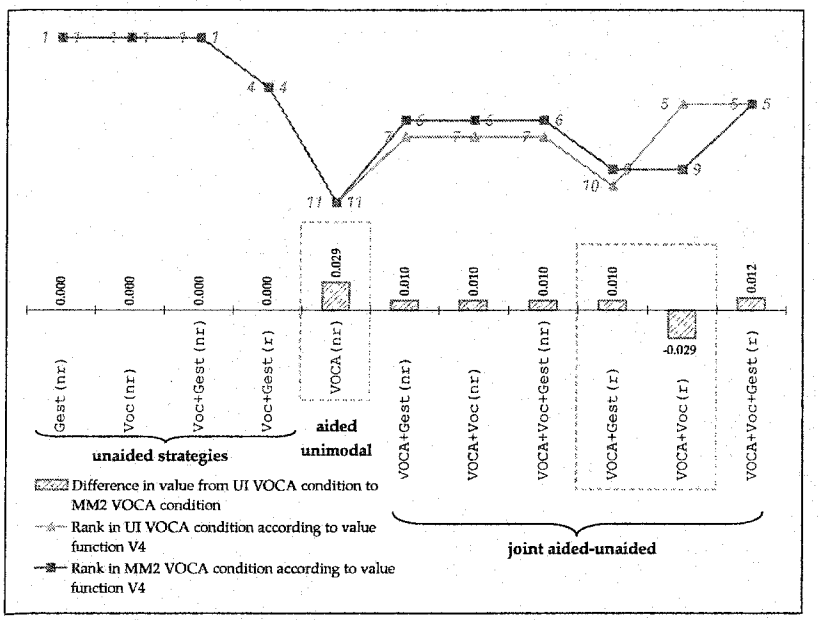
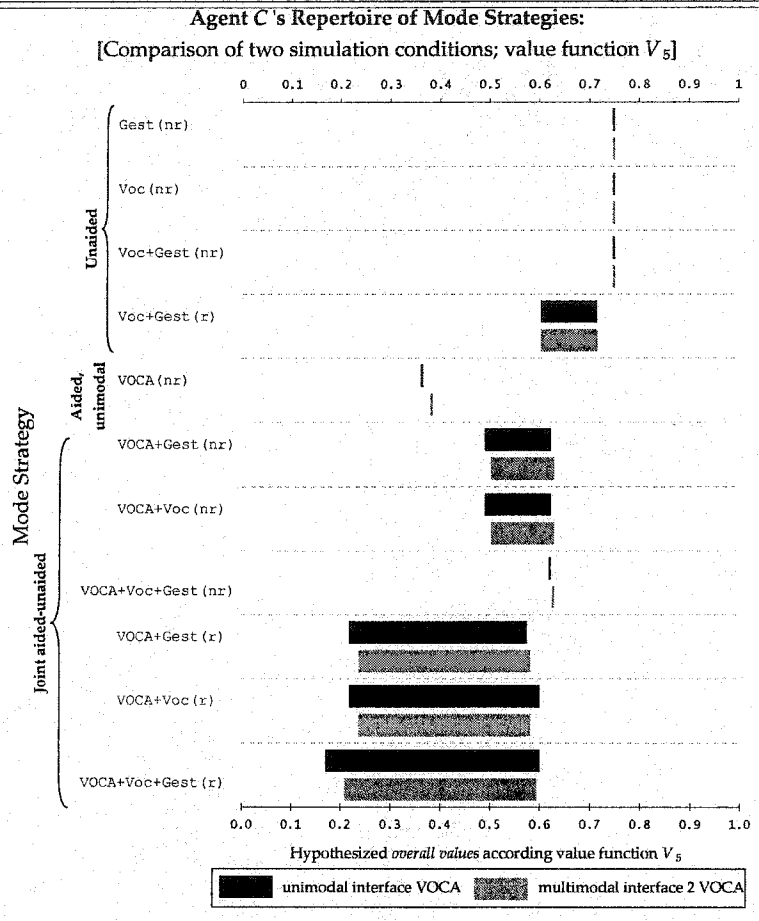


Figure A.10 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* and *multimodal interface VOCA* conditions when the value function V_5 is used (top graph) and the respective rankings that were obtained by MSIM (using the maximum overall value of each mode strategy) (bottom diagram).



Ranking of Agent *C* 's Mode Strategies:

A comparison of the *unimodal VOCA* and *multimodal VOCA* conditions

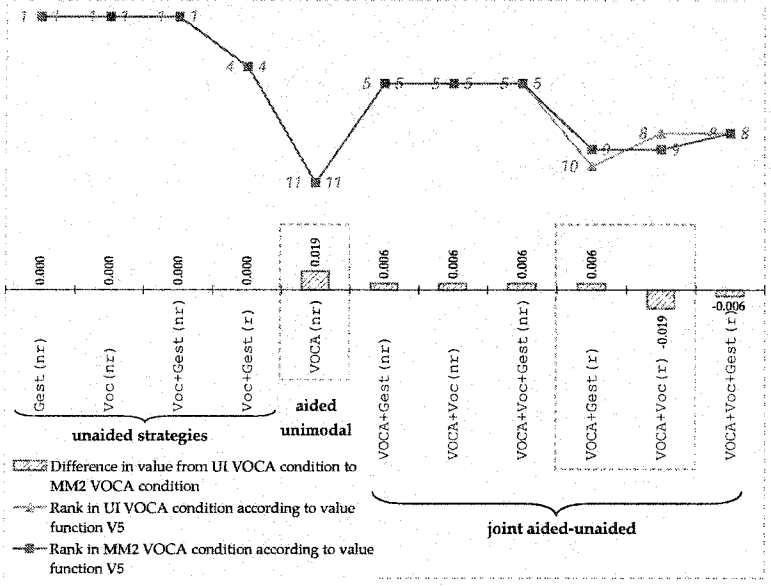
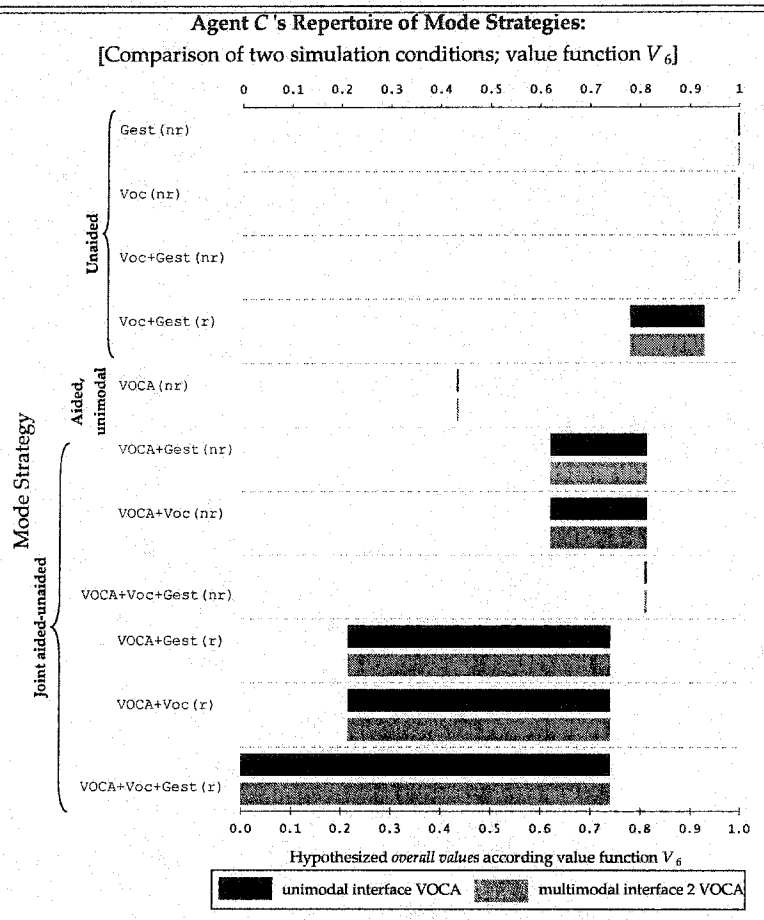
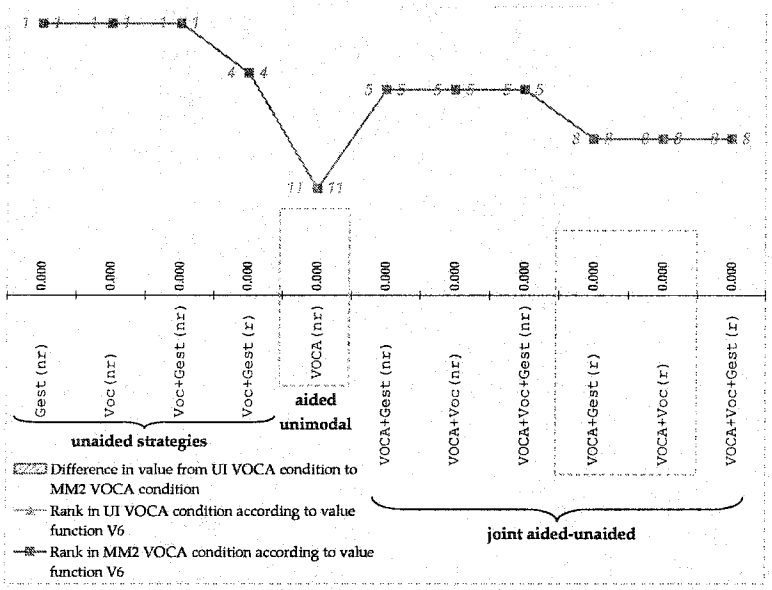


Figure A.11 Profiles of agent *C*'s mode strategies in the *unimodal interface VOCA* and *multimodal interface VOCA* conditions when the value function V_6 is used (top graph) and the respective rankings that were obtained by MSIM (using the maximum overall value of each mode strategy) (bottom diagram).





A.6 Glossary

AAC system user: an individual who makes use of an AAC system. Since both individuals in a dyad make use of communicative strategies, and since AAC systems include these as one of their system components, both individuals in a dyad are, technically speaking, AAC system users. However, in the AAC literature, this term is often used to refer to the aided communicator.

Aided communicator: An interlocutor who uses a communication aid (as warranted by the communication partner and the communicative scenario). The repertoire of mode strategies that is available to an aided communicator includes the unaided, aided-unimodal, and joint unaided-aided mode strategies.

Aided dyad: A pair of interlocutors in which one interlocutor uses a communication aid (because of a communication disorder) and the other does not. The communicative exchange is mediated by an AAC system (since a communication aid implies the use of an AAC system). Two variants can be distinguished on the basis of the familiarity of interlocutors: familiar-aided and unfamiliar-aided.

Aided mode: The mode of synthesized speech, which is afforded by a VOCA (one type of communication aid).

***Aided-unimodal* strategy:** A manner in which an interlocutor performs a communicative action using only synthesized speech.

Interlocutor: Similar to *participant* (an individual who is engaged in a collaborative task), although this term emphasizes individual's participation in the communicative process. In some contexts, the term *speaker* is used interchangeably with *interlocutor*. But since a variety of behaviours other than acts of speaking can be communicative, the more-general term *interlocutor* will be used.

***Joint aided-unaided* mode strategy:** A manner in which an interlocutor performs a communicative action using both the aided mode of articulation and one or more unaided modes of articulation.

Mode: A manner in which some action can be performed, carried out, or conducted (where the action depends on the context of the term's use). For example, a mode of articulation, a mode of sensory input, a mode of communication.

Mode strategy: A manner in which an interlocutor can employ his or her modes of articulation to perform a communicative action.

Multimodal articulation: A component sub-process of multimodal utterance design whereby an underlying communicative plan is articulated by an interlocutor, using his or her communicative resources.

Multimodal strategy: A manner in which an interlocutor performs a communicative action using two or more modes of articulation.

Multimodal utterance design: A term used by Clark [1996] to refer to the process whereby interlocutors derive and perform their communicative actions. The process entails the derivation of a communicative plan, the derivation of a multimodal surface realization for it, and the performance of the multimodal surface realization.

Multimodal VOCA: A VOCA that has an interface that is capable of recognizing input actions that are produced by more than one input mode (such as vocalizations, facial expressions, or gestures of the hand or head). The interface interprets each action as an independent unit (such as the system described by Treviranus et al. [1991]; see page 36) or as one component of a larger composite (such as the system described by Roy et al. [1993a]; see page 39).

Participant: An individual who is engaged in a collaborative task.

Surface Realization: In essense, the observable portion of a multimodal communicative action. Predicated on a view in which communicative actions, or utterances, are defined as a process: "an utterance is ... a process that has an internal development and has ... surface linguistic constituents [in] its final stage" [McNeill, 1992, p. 218]. Aided communicators often incorporate synthesized speech in the surface realizations that they produce. See section 6.2.1 more a detailed discussion.

Unaided communicator: An interlocutor who does not use a communication aid.

Unaided dyad: A pair of interlocutors in which neither interlocutor has a communication disorder and neither uses a communication aid. Thus, the communicative exchange is not mediated by an AAC system.

Unaided mode: Any mode of articulation that is available to an interlocutor using his or her own communicative effectors and does not entail the use of a communication aid.

***Unaided* mode strategy:** A manner in which an interlocutor performs a communicative action using only unaided modes of articulation.

Unimodal strategy: A manner in which an interlocutor performs a communicative action using a single mode of articulation.

Unimodal VOCA: A VOCA that has an interface that is capable of recognizing the input actions that are produced using a single mode of input.

VOCA: Stands for Voice-Output Communication Aid, a particular type of communication aid.

Bibliography

James Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, 2nd edition, 1994.

Norman Alm, John L. Arnott, and Alan F. Newell. Evaluation of a text-based communication system for increasing conversational participation and control. In *Proceedings of RESNA '92 15^{th} Annual Conference*, pages 366–368. RESNA Press, 1992a.

Norman Alm, John L. Arnott, and Alan F. Newell. Prediction and conversational momentum in an augmentative communication system. *Communications of the ACM*, 35(5):47–57, 1992b.

Norman Alm, Mark Nichol, and John L. Arnott. The application of fuzzy set theory to the storage and retrieval of conversational texts in an augmentative communication system. In *Proceedings of the Rehabilitation Engineering Society of North America (RESNA)*, pages 127–129, 1993.

J. L. Arnott, N. Alm, and A. F. Newell. A text database as a communication prosthesis. In *Choice For All: Proceedings of the International Conference of the Association for the Advancement of Rehabilitation Technology (ICAART)*, pages 76–77, Montreal, 1988. RESNA.

American Speech-Language-Hearing Association ASHA. Competencies for speech-language pathologists providing services in augmentative communication. *ASHA*, 31(3):107–110, 1989.

American Speech-Language-Hearing Association ASHA. Report: Augmentative and alternative communication. *ASHA*, 33(3):9–12, 1991. Supplementary No. 5.

J. L. Austin. *How to Do Things with Words*. Harvard University Press, 1962.

B. R. Baker. Minspeak: A semantic compaction system that makes self-expression easier for communicatively disabled individuals. *Byte*, 7(9):186–202, 1982.

Melanie Baljko. The computational simulation of multimodal, face-to-face communication constrained by physical disabilities. In *Proceedings of ESSLLI 2000 Workshop Integrating Information from Different Channels in Multi-Media Contexts*, pages 1–10, Birmingham, UK, August 6–10 2000a. European Association for Logic, Language, and Information.

Melanie Baljko. Incorporating multimodality in the design of interventions for communication disorders. In Patric Dahlqvist, editor, *Proceedings of 4th SSoMC, the Fourth Swedish Symposium on Multimodal Communication*, pages 13–14. Stockholm University/KTH, October 26–27 2000b.

Melanie Baljko. Articulatory adaptation in multimodal communicative action. In Cindi Thompson, Tim Paek, and Eric Horvitz, editors, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) Workshop on Adaptation in Dialogue Systems*. Pittsburgh, PA, June 2001a.

Melanie Baljko. The evaluation of microplanning and surface realization in the generation of multimodal acts of communication. In Catherine Pelachaud and Isabella Poggi, editors, *Proceedings of the Workshop on Multimodal Communication and Context in Embodied Agents, Fifth International Conference on Autonomous Agents*, pages 89–94. Montreal, Canada, May 2001b.

Jan L. Bedrosian, Linda A. Hoag, Stephen N. Calculator, and Barry Molineaux. Variables influencing perceptions of the communicative competence of an adult augmentative and alternative communication system user. *Journal of Speech and Hearing Research*, 35(5):1105–1113, October 1992.

K. L. Berge. Communication. In R. E. Asher, editor, *The Encyclopedia of Language and Linguistics*, volume 2, pages 614–620. Pergamon Press, Great Britain, 1994.

L. E. Berstein. *The Vocally Impaired: Clinical Practice and Research*. Grune & Stratton, 1988.

D. Beukelman and K. Yorkston. Non-vocal communication — Performance evaluation. *Archives of Physical Medicine and Rehabilitation*, 61:272–275, 1980.

David R. Beukelman and Pat Mirenda. *Augmentative and alternative communication: management of severe communication disorders in children and adults*. Paul H. Brookes, Baltimore, MD, second edition, 1998.

S. Blackstone and E. Cassatt. Communicative competence in communication aid users and their partners. Paper presented at the Third International Conference on Augmentative and Alternative Communication, Boston, MA, 1984.

S. Blackstone and H. Pressman. *"Outcomes in AACČonference Report: Alliance '95*. Augmentative Communication, Monterey, CA, 1995.

Doreen M. Blischak and Lyle L. Lloyd. Multimodal augmentative and alternative communication: Case study. *Augmentative and Alternative Communication*, 12(1):37–46, March 1996.

Richard Bolt. *The Human Interface: Where People and Computers Meet*. Lifelong Learning Publications, Belmont, CA, 1984.

Dennis Bouchard. Sign language and language universals: The status of order and position in grammar. *Sign Language Studies*, 91(2):101–159, 1996.

Serge Brédart, Tim Brennan, and Tim Valentine. Dissociations between the processing of proper and common names. *Cognitive Neuropsychology*, 14(2):209–217, 1997.

K. Bühler. The deictic field of language and deictic words. In R. J. Jarvella and W. Klein, editors, *Speech, Place, and Action: Studies in Deixis and Related Topics*, pages 9–30. John Wiley & Sons, London, UK, 1982. An adbridged English version of Part 2 (chapters 7 and 8) of Bühler (1934), prepared by Jarvella and Klein.

Harry Bunt, René Ahn, Robbert-Jan Beun, Tijn Borghuis, and Kees van Overveld. Cooperative multimodal communication in the DenK project. In *Proceedings of the International Conference on Cooperative Multimodal Communication CMC/95*, pages 79–102, Eindhoven, May 1995.

D. Byrd and E. Saltzman. Speech production. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1072–1076. MIT Press, Cambridge, MA, second edition, 2002.

G. A. Calvert, M. J. Brammer, and S. D. Iversen. Crossmodal identification. *Trends In Cognitive Sciences*, 2(7):247–253, July 1998.

J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, and H. Yan. Human conversation as a system framework: Designing embodied conversational agents. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, chapter 2, pages 29–63. MIT Press, Cambridge, MA, 2000.

J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, M. Achorn, T. Beckett, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics, SIGGRAPH'94*, pages 413–420, Orlando, FL, 1994.

Justine Cassell and Matthew Stone. Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems. In Susan Brennan, Alain Giboin, and David Traum, editors, *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 34–42, North Falmouth, Mass, November 1999.

Noam Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.

H. H. Clark. *Arenas of Language Use*. The University of Chicago Press and the Center for the Study of Language and Information, 1992.

H. H. Clark. *Using Language*. Cambridge University Press, 1996.

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

Robert V. Conti, Jeffrey Micher, and Gail VanTatenhove. Frequently asked questions about Minspeak. WWW, March 1998. http://kaddath.mt.cs.cmu.edu/scs/faq.htm.

Ann Copestake. Applying natural language processing techniques to speech prostheses. In *Proceedings of the AAAI Fall Symposium on developing assistive techniques for people with disabilities*, MIT, Cambridge, MA, 1996.

Robert T. Craig. Communication Theory as a field. *Communication Theory*, 9(2):119–161, May 1999.

Robert Dale. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, University of British Columbia, Vancouver, June 26-29 1989.

R. I. Damper. Text composition by the physically disabled: A rate prediction model for scanning input. *Applied Ergonomics*, 15(4):289–296, December 1984.

R. B. Dannenberg and M. Blattner. Introduction: The trend toward multimedia interfaces. In M. Blattner and R. B. Dannenberg, editors, *Multimedia Interface Design*, pages xvii–xxv. ACM Press, New York, 1992.

Status of Disabled Persons Secretariat, Department of the Secretary of State of Canada DSS-C. A way with words: Guidelines and appropriate terminology for the portrayal of persons with disabilities, 1991.

John Dunaway, Patrik Demasco, Denise Peischl, and Alice Smith. A pilot study for multimodal input in computer access. In *Proceedings of the RESNA '96 Annual Conference*, pages 319–321. RESNA Press, 1986.

J. S. Dyer and R. K. Sarin. Measurable multiattribute value functions. *Operations Research*, 27(4): 810–822, 1979.

A. D. N. Edwards. Multimodal interaction and people with disabilities. In Björn Granström, David House, and Inger Karlsson, editors, *Multimodality in Language and Speech Systems*, pages 73–92. Kluwer, Dordrecht, The Netherlands, 2002.

David Efron. *Gesture, Race and Culture*. Mouton, The Hague, 1972. Originally published under title: Gesture and Environment, New York, King's Crown Press, 1941.

Paul Ekman and Wallace V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System: Investigator's Guide*. A Human Face, Salt Lake City, UT, 2002 revised edition edition, 2002.

Paul Ekman and Wallane V. Friesen. The repetoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

Richard A. Foulds. Communication rates for nonspeech expression as a function of manual tasks and linguistic constraints. In *Proceedings of the International Conference on Rehabilitation Engineering (ICRE-80)*, pages 83–87, Toronto, Canada, June 1980.

K. L. Garrett and D. R. Beukelman. Adults with severe aphasia. In David R. Beukelman and Pat Mirenda, editors, *Augmentative and alternative communication: management of severe communication disorders in children and adults*, pages 465–499. Paul H. Brookes, Baltimore, MD, second edition, 1998.

Sharon L. Glennen and Denise C. DeCoste. *The handbook of augmentative and alternative communication*. Singular Publishing Group, 1997.

Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Language, thought, and culture. Academic Press, New York, 1981.

Philip Babcock Gove, editor. *Webster's Seventh New Collegiate Dictionary*. G & C Merriam Company, Springfield, MA, 1967.

H.P. Grice. Meaning. *Philosophical Review*, 66:377–388, 1957. Reprinted in D. Steinberg and L. Jakobovits, *Semantics: An interdisciplinary reader in Philosophy, Linguistics and Psychology*, Cambridge University Press, 1972.

H.P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics, Speech Acts*, volume 3, pages 41–58. Academic Press, 1975.

P. Griffith. Mode-switching and mode-finding in a hearing child of deaf parents. *Sign Language Studies*, 48:195–221, 1985.

Y. Gu. The impasse of perlocution. *Journal of Pragmatics*, 20:405–432, 1993.

Marianne Gullberg. Gestures in spatial descriptions. In *Working Papers of the Department of Linguistics and Phonetics, Lund University*, volume 47, pages 87–98. Lund University, 1999. Available at `http://www.ling.lu.se/disseminations/wp_overview.html`, also `http://www.ling.lu.se/disseminations/pdf/47/Gullberg.pdf`.

Peter A. Heeman and Graeme Hirst. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382, 1995.

E. Horvitz and T. Paek. A computational architecture for conversation. In *Proceedings of the Seventh International Conference on User Modeling*, pages 201–210, Banff, Canada, June 1999. New York: Springer Wien.

E. Horvitz and T. Paek. Deeplistener: Harnessing expected utility to guide clarification dialog in spoken language systems. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, November 2000.

M. Huer. White's gestural system of the lower extremities. *Communicating Together*, 5:3–4, 1987.

Katherine C. Hustad, Ray D. Kent, and David R. Beukelman. DECTalk and MacinTalk speech synthesizers: Intelligibility differences for three listener groups. *Journal of Speech, Language, and Hearing Research*, 41(4):744–752, August 1998.

G. Jones, F. E. Ritter, and D. J. Wood. Using a cognitive architecture to examine what develops. *Psychological Science*, 11(2):93–100, 2000.

Ashish Kapoor and Rosalind W. Picard. A real-time head nod and shake detector. In *Proceedings from the Workshop on Perspective User Interfaces PUI'01*, pages ??–??, Orlando, Florida, November 2001.

Simeon Keates and Peter Robinson. The use of gestures in multimodal input. In *Proceedings of the Third International ACM Conference on Assistive Technologies*, pages 35–42, Marina del Rey, CA, April 15–17 1998.

A. Kendon. How gestures can become like words. In Fernando Poyatos, editor, *Cross-Cultural Perspectives in Non-Verabl Communication*, pages 131–141. C. J. Hogrefe, Toronto, 1988.

David E. Kieras and David E. Meyer. An overview of the EPIC architecture for cognition and performance with application to human–computer interaction. *Human–Computer Interaction*, 12: 391–438, 1997.

Michael Kipp. From human gesture to synthetic action. In Catherine Pelachaud and Isabella Poggi, editors, *Proceedings of the Workshop on Multimodal Communicatiom and Context in Embodied Agents, Fifth International Conference on Autonomous Agents*, pages 9–14. Montreal, Canada, May 2001.

R. M. Krauss and S. Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114, 1964.

R. M. Krauss and S. Weinheimer. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personaility and Social Psychology*, 14:343–346, 1966.

R. M. Krauss and S. Weinheimer. Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6:359–363, 1967.

Joanne P. Lasker and Jan L. Bedrosian. Promoting acceptance of augmentative and alternative communication by adults with acquired communication disorders. *Augmentative and Alternative Communication*, 17(3):141–153, September 2001.

J. Light. Interaction involving individuals using augmentative and alternative communication: State of the art and future directions. *Augmentative and Alternative Communication*, 4:66–82, 1988.

J. Light, B. Collier, and P. Parnes. Communicative interaction between young nonspeaking physically disabled children and their primary caregivers: Part I – Discourse patterns. *Augmentative and Alternative Communication*, 1(3):74–83, 1985a.

J. Light, B. Collier, and P. Parnes. Communicative interaction between young nonspeaking physically disabled children and their primary caregivers: Part II – communicative function. *Augmentative and Alternative Communication*, 1(4):98–107, 1985b.

J. Light, B. Collier, and P. Parnes. Communicative interaction between young nonspeaking physically disabled children and their primary caregivers: Part III – Modes of communication. *Augmentative and Alternative Communication*, 1(4):125–133, 1985c.

J. Light, P. Siegel, and P. Parness. The effect of message encoding techniques on recall by literate adults using AAC systems. *Augmentative and Alternative Communication*, 6:184–197, 1990.

Janice C. Light. *Message Encoding Techniques for Augmentative Communication Systems: An Investigation of the Recall Performances of Nonspeaking Physically Disabled Adults*. PhD thesis, University of Toronto, 1989.

Janice C. Light. Do augmentative and alternative communication interventions really make a difference?: The challenges of efficacy research. *Augmentative and Alternative Communication*, 15(1): 13–24, March 1999.

Lyle L. Lloyd, Donald R. Fuller, and Helen H. Arvidson. *Augmentative and Alternative Communication: a handbook of principles and practices*. Allyn and Bacon, Needham Heights, MA, 1997.

Daniel Marcu. Perlocutions: The Achilles' Heel of Speech Act Theory. *Journal of Pragmatics*, 32(12): 1719–1741, 2000.

Jean-Claude Martin and Dominque Béroule. Temporal codes within a typology of cooperation between modalities. *Artificial Intelligence Review*, 9(2–3):95–102, 1995.

Jean-Claude Martin, Sarah Grimard, and Katerina Alexandri. On the annotation of multimodal behavior and computation of cooperation between modalities. In Catherine Pelachaud and Isabella Poggi, editors, *Proceedings of the Workshop on Multimodal Communicatiom and Context in Embodied Agents, Fifth International Conference on Autonomous Agents*, pages 1–7. Montreal, Canada, May 2001.

Angus McIntyre. Babel: A testbed for research in origins of language. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Annual International Conference on Computational Linguistics (COLING-ACL '98)*, pages 830–835, August 1998.

David McNeill. *Hand and mind: what gestures reveal about thought*. University of Chicago Press, 1992.

P. Mirenda and P. Mathy-Laikko. Augmentative and alternative applications for persons with severe congenital disorders: An introduction. *Augmentative and Alternative Communication*, 5(1): 3–13, 1989.

Melody M. Moore and Philip R. Kennedy. Human factors issues in the neural signals direct brain-computer interfaces. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies (ASSETS)*, pages 114–120, Arlington, VA, November 13–15 2000.

Clifford Nass, Katherine Isbister, and Eun-Ju Lee. Truth is beauty: Researching embodied conversational agents. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, chapter 13, pages 374–402. MIT Press, Cambridge, MA, 2000.

National Center on Disability and Journalism NCDJ. Style guide of the National Center on Disability and Journalism, 2002.

Laurence Nigay and Joëlle Coutaz. A design space for multimodal interfaces: concurrent processing and data fusion. In *Proceedings of the INTERCHI'93 Conference on Human Factors in Computing Systems*, pages 172–179. ACM Press, 24–29 April 1993.

Tim Paek and Eric Horvitz. Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational system. In *AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, Cape Cod, MA, November 5-7 1999.

Tim Paek and Eric Horvitz. Conversation as action under uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pages 455–464, Stanford, CA, June 2000.

Isabella Poggi and Catherine Pelachaud. Performative faces. *Speech Communication*, 26(1–2):5–21, October 1998.

R. Power. The organization of purposeful dialogs. *Linguistics*, 17(1/2):105–152, 1979.

F. Poyatos. Interaction functions and limitations of verbal and nonverbal behaviors in natural conversation. *Semiotica*, 30(3–4):211–244, 1980.

F. Quek, R. Bryll, D. McNeill, and M. Harper. Gestural origo and loci-transitions in natural discourse segmentation. In *Proceedings of IEEE Workshop on Cues in Communication*, 9 December 2001.

M. J. Reddy. The conduit metaphor — a case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*. Cambridge University Press, Cambridge, 1979.

Frank E. Ritter, Gordon D. Baxter, Gary Jones, and Richard M. Young. Supporting cognitive models as users. *ACM Transactions on Computer-Human Interaction*, 7(2):141–173, June 2000.

D. Roy, M. Panayi, R. Foulds, R. Erenshteyn, W. Harwin, and R. Fawcus. The enhancement of interaction for people with severe speech and physical impairment through the computer recognition of gesture and manipulation. *Presence: Teleoperators and Virtual Enviroments*, 3(3):227–235, Summer 1994a.

D. Roy, M. Panayi, R. Foulds, R. Fawcus, R. Erenshteyn, and W. Harwin. Computer recognition of gestures for augumentative and alternative communication. In *Proceedings 6th Biennial Conference of the International Society for Augmentative and Alternative Communication (ISAAC'94)*, Maastricht, Netherlands, October 1994b.

D. M. Roy. Computer recognition of movement for people with athetoid cerebral palsy. In *Proceedings 5th Biennial Conference of the International Society for Augmentative and Alternative Communication (ISAAC'92)*, page 164, 1992.

David M. Roy, Marilyn Panayi, Roman Erenshteyn, Richard Foulds, and Robert Fawcus. Gestural human-machine interaction for people with severe speech and motor impairment due to cerebral palsy. In *CHI'94 Conference Companion on Human Factors in Computing Systems*, pages 313–314. ACM Press, 1994c. ISBN 0-89791-651-4. doi: http://doi.acm.org/10.1145/259963.260375.

David M. Roy, Marilyn Panayi, William S. Harwin, and Robert Fawcus. Advanced input methods for people with cerebral palsy: A vision of the future. In *Proceedings of the Rehabilitation Engineering Society of North America*, pages 99–101, 1993a.

David M. Roy, Marilyn Panayi, William S. Harwin, and Robert Fawcus. The enhancement of interaction for people with severe speech and physical impairment through the computer recognition of gesture and manipulation. In *Proceedings of the CSUN Conference on Virtual Reality and Persons with Disabilities*, 1993b.

The Research & Training Center on Independent Living RTC/IL. Guidelines for reporting and writing about people with disabilities, 2001. Sixth Edition.

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

N. Schiavetti and D. Metz. *Evaluating Research in Communicative Disorders*. Allyn & Bacon, Boston, third edition, 1997.

Deborah Schiffrin. *Approaches to Discourse*. Blackwell, Oxford, UK, 1994.

Ralf W. Schlosser and Ursula Braun. Efficacy of AAC interventions: Methodologic issues in evaluating behavior change, generalization, and effects. *Augmentative and Alternative Communication*, 10(4):207–223, December 1994.

Ralf W. Schlosser and David L. Lee. Promoting generalization and maintenance in augmentative and alternative communication: A meta-analysis of 20 years of effectiveness research. *Augmentative and Alternative Communication*, 16(4):208–226, December 2000.

J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, London, 1969.

J. R. Searle. *Expression and Meaning*. Cambridge University Press, 1979.

George H. Shames, Elisabeth H. Wiig, and Wayne A. Secord, editors. *Human Communication Disorders: An Introduction*. Allyn and Bacon, Needham Heights, MA, fifth edition, 1998.

Fraser Shein, Nicholas Brownlow, Jutta Treviranus, and Penny Parnes. Climbing out of the rut: The future of interface technology. In B. Mineo, editor, *Proceedings of the Visions Conference: Augmentative and Alternative Communication in the Next Decade*, Univerisity of Delaware/Alfred I. duPont Institute, Wilmington, DE, 1990. Applied Science and Engineering Laboratories, Alfred I. duPont Institute.

Fraser Shein, Tom Nantais, Rose Nishiyama, Cynthis Tam, and Paul Marshall. Word cueing for persons with writing difficulties: WordQ. In *Proceedings of the CSUN 16th Annual Conference on Technology for Persons with Disabilities*, 2001.

J. A. Simpson and E. S. C. Weiner, editors. *The Oxford English Dictionary*. Oxford University Press, second edition, 1989.

A. Smith, J. Dunaway, P. Demasco, and D. Peichl. Multimodal input for computer access and alternative communication. In *Proceedings of the Second Annual ACM Conference on Assistive Technologies ASSETS'96*, pages 80–85, Vancouver, Canada, 1996. ACM Press. doi: http://doi.acm.org/10.1145/228347.228361.

Matthew Stone. Representing communicative intentions in collaborative conversational agents. In *Proceedings of AAAI Fall Symposium on Intent Inference for Collaborative Tasks*, 2001.

Carol S. Swindell, Audrey L. Holland, and O. M. Reinmuth. Aphasia and related adult disorders. In George H. Shames, Elisabeth H. Wiig, and Wayne A. Secord, editors, *Human Communication Disorders: An Introduction*, pages 472–509. Allyn and Bacon, Needham Heights, MA, fifth edition, 1998.

Kristinn Rúnar Thórisson. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. PhD thesis, Massachusetts Institute of Technology, 1996.

John Todman and Norman Alm. Pragmatics and AAC approaches to conversational goals. In Ann Copestake, Stefan Langer, and Sira Palazuelos-Cagigas, editors, *Natural Language Processing for Communication Aids, Proceedings of a Workshop Sponsored by the Association for Computational Linguistics*, pages 1–8, Madrid, Spain, July 1997.

John Todman and Norman Alm. Modelling conversational pragmatics in communication aids. *Journal of Pragmatics*, 35(4):523–538, April 2003.

D. R. Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence (Special Issue on Non-literal Language)*, 8(3), 1992.

David Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, December 1994. Also published as technical report 545.

David R. Traum. Speech acts for dialogue agents. In Michael Wooldridge and Anand Rao, editors, *Foundations And Theories Of Rational Agents*, pages 169–201. Kluwer Academic Publishers, 1999.

David R. Traum and James F. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 1994.

J. Treviranus, F. Shein, S. Haataja, P. Parnes, and M. Milner. Speech recognition to enhance computer access for children and young adults who are functionally nonspeaking. In J. J. Presperin, editor, *Proceedings of the 14th Annual Conference of the Rehabilitation Engineering Society of North America (RESNA'91)*, pages 308–310, Kansas City, MO, 1991. RESNA.

V. Turchin, C. Joslyn, and F. Heylighen (editors). Principia Cybernetica Web. Published on line at: `http://pespmc1.vub.ac.be`, 1991–2003.

Pascal Vaillant and Michael Checler. Intelligent voice prosthesis: converting icons into natural language sentences. In *Actes de "Montpellier'95 - 4èmes Journées Internationales 'L'Interface des Mondes Rels et Virtuels' "*, 26 – 30 June 1995.

Ielka van der Sluis. An empirically motivated algorithm for the generation of multimodal referring expressions. In *Proceedings of the Student Research Workshop, the 39th Annual Meeting of the Association for Computational Linguistics*, pages 67–72, Toulouse, France, July 6–11 2001.

Ielka van der Sluis and Emiel Krahmer. Generating referring expressions in a multimodal context. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting*, number 37 in Language and Computers: Studies in Practical Linguistics, pages 158–176. Rodopi, Amsterdam, 2000.

J. A. Van Dyke. Word prediction for disabled users: Applying natural language processing to enhance communication. Master's thesis, University of Delaware, Newark, DE, 1991. Thesis for Honors Bachelor of Arts in Cognitive Studies.

G. C. Vanderheiden. Overview of the basic selection techniques for augmentative communication: Present and future. In L. E. Berstein, editor, *The Vocally Impaired: Clinical Practice and Research*, pages 5–39. Grune & Stratton, Philadelphia, 1988.

Horabail S. Venkatagiri. Efficient keyboard layouts for sequential access in augmentative and alternative communication. *Augmentative and Alternative Communication*, 15(2):126–134, June 1998.

Anne Warrick. Sociocommunicative considerations within augmentative and alternative communication. *Augmentative and Alternative Communication*, 4(1):45–51, 1988.

David E. Yoder and Arlene Kraat. Intervention issues in nonspeech communication. In Jon Miller, David E. Yoder, and Richard Schiefelbusch, editors, *Contemporary Issues in Language Intervention*, ASHA Reports 12, chapter 3, pages 27–51. American Speech-Language-Hearing Association, Rockville, Maryland, 1983.

Citation Index

Reddy [1979], 14
Ritter et al. [2000], 133, 134
Roy [1992], 2
Roy et al. [1993b], 2, 39
Roy et al. [1993a], 2, 37, 39
Roy et al. [1994a], 122
Roy et al. [1994b], 2
Roy et al. [1994c], 2, 39
RTC/IL [2001], 26, 27
Sacks et al. [1974], 15
Schiavetti and Metz [1997], 31
Schiffrin [1994], 50
Schlosser and Lee [2000], 31, 32
Schlosser and Braun [1994], 31
Searle [1969], 3, 14, 18
Searle [1979], 18
Shames et al. [1998], 28
Shein et al. [2001], 35
Shein et al. [1990], 2, 38, 39, 55, 122, 126
Smith et al. [1996], 2, 40
Swindell et al. [1998], 13
Thórisson [1996], 50, 51, 140, 141
Todman and Alm [2003], 35
Todman and Alm [1997], 15
Traum and Hinkelman [1992], 3, 14
Traum [1994], 20
Traum and Allen [1994], 70
Traum [1999], 136
Treviranus et al. [1991], 2, 36, 98
Vaillant and Checler [1995], 15
Vanderheiden [1988], 36
van der Sluis and Krahmer [2000], 71
van der Sluis [2001], 71
Venkatagiri [1998], 36
Warrick [1988], 55
Yoder and Kraat [1983], 37

Subject Index