Paper 1

The Computational Simulation of Multimodal, Face-to-Face Communication Constrained by Physical Disabilities

Melanie Baljko
University of Toronto

1 Introduction

In face-to-face interaction, interlocutors often use several modes of articulation simultaneously. An interlocutor's communication will often be multimodal even when he or she knows the other interlocutors cannot perceive all of the modes of communication (e.g., people often gesture while speaking on the telephone). Our present inquiry — which incorporates computational modeling in conjunction with the analysis of, and comparison to empirical data — is motivated by the desire to understand a particular design space (to be described below) and is relevant to other research that seeks to understand these "complex signals" in human-human and human-computer interaction.

The articulation and perception of multimodal communicative acts always take place in an **articulatory-perceptual (A-P) channel**, the primary ones being **auditory-oral** and the **visual-gestural** channels. In some cases, multimodal communicative behaviour involves multiple channels. For instance, speech has both oral and gestural properties, and, in face-to-face conversation, is perceived both auditorially and visually. But, as in the case of a telephone conversation, the mode of speech might involve a single channel only. Single-channel communication can still be multimodal, however. For instance, a face-to-face conversation between two interlocutors using a signed language will take place solely within the visual-gestural channel, even though the modes of speech, gesture, and eye gaze are all used.

Even though the notion of channel and mode are intertwined, their definitions must be separated. A **mode** of communication is not only the means by which communicative behaviour can be expressed, but also, to some degree, the abstract properties of the shared system of meaning in which the communicative act is interpreted (e.g., a distinction is often made between the modes of speech and vocalization). The production and perception of communicative behaviour is dependent on the functioning of each interlocutor's **sensorimotor articulatory and perceptual apparatus** — the neuromuscular **articulators** for expressive behaviour, such as the resonance or phonatory articulators used for the modes of speech and vocalization, the ocumotor articulators used for eye gaze, or the muscoloskeletal articulators used for gesture; and the physiological apparatus for the sensory perception.

The functioning of an interlocutor's sensorimotor apparatus is one dimension along which the production of *multimodal communicative acts*[1] can be analyzed. This view contrasts with the approach taken by others in exploring the dimension of *semantic or pragmatic content* with respect to the production (Cassell et al., 1994) or to the comprehension or interpretation (Johnston, 1998; McGee et al., 1998; Oviatt, 1999) of multimodal communicative acts.

[1]The term *composite signal* is intuitively appealing for referring to instances of multimodal communicative

Proceedings of the Workshop on Integrating Information from Different Channels in Multi-Media-Contexts at ESSLLI MM
K. Holmqvist, P. Kühnlein & H. Rieser (eds).

2 Motivation

In certain populations of communicators, the functioning of the sensorimotor apparatus is constrained. The modes of speech, gesture, facial expression, and (rarely) eye gaze, may each be partially or completely unusable due to congenital or acquired *physical disorder* (for example, dysarthria caused by cerebral palsy or amyotrophic lateral sclerosis). In many cases, *linguistic* communication is not possible due to inadequate function of the modes of speech (for spoken language) or gesture (for signed language). Because these modes support linguistic communication, they are particularly salient. The clinical research area of **augmentative and alternative communication** attempts to develop useful interventions in these cases, and often these interventions come in the form of **communication devices**. Many such devices are beneficial because they can be used to access another — albeit aided — mode: synthesized speech.

Unfortunately, the design of these devices does not adequately account for a number of issues:

- Effective communication is best accomplished by the use of a *repertoire* of modes working in concert together — including both the aided mode and the "native" (unaided) modes.
- Certain modes are particularly *salient* for particular discourse functions. For example, the mode of eye gaze often figures prominently in turn-taking. Since the use of particular discourse functions varies over the course of a communicative exchange, the *relative saliency* of modes can vary over time too.
- It is suboptimal for the device interface to require the use of certain modes at certain times (since, in many cases, they are already being used, the device is essentially competing with other communicative modes).
- The aided mode needs to provide different communicative functions at different times during an interaction (such as to provide back-channel feedback, to establish common ground, to signal misunderstanding, to initiate repair, to pass and hold the turn, and so on).

Any redesign of these communication devices will require a theoretical foundation for making design decisions about the relative saliency of modes for constrained communicators. To this end, we have developed a model of constrained, face-to-face multimodal communication that describes the relative saliency of the modes at particular times within communicative scenarios. The computational implementation of this model is currently iterating between the stages of evaluation and modification. This model has potential applications to other areas with analogous populations of constrained communicators — for example, the users of Web-based interfaces when the availability of network bandwidth places constraints on the functioning of intelligent multimedia presentation managers (Arens and Hovy, 1995); the users of interfaces in contexts in which they might become functionally disabled due to constraints arising from noise or fatigue (Newell et al., 1995); and users of video-conferencing systems where the availability of resources for transmitting information is constrained.

3 Simulating Communicative Agents

In this research, we hypothesize that certain aspects of communicative behaviour can be modeled as the process of finding solutions to *constraint satisfaction problems*. In particular, in certain situations AAC-system users use modes that are "risky" yet convenient (for instance, with familiar communication partners), while in other situations, they use modes that are "redundant" but fatiguing (e.g., with unfamiliar communication partners).

The idea that modes can be characterized in terms of their *cost of articulation*, and *likelihood of causing misunderstanding* is useful for characterizing not only AAC-based communication but also

behaviour, but it carries the implication of a message-passing model of communication. In its place, the term *multimodal communicative act* will be used; this is a generalization of the term *locutionary act* and we believe it to be a reasonable extension to the term as Searle and Austin intended (Austin, 1962; Searle, 1979). For a discussion of the issues arising from the definition of a *multimodal* speech act, see Baljko (forthcoming).

more-generalized instances of face-to-face communication. We hypothesize that what is intuitively described as *synergy* in multimodal communication can be formally defined in terms of particular solution characteristics, which arise under certain circumstances. In fact, a recurrent theme in this line of research is that AAC-based communication is simply another variant of "typical" face-to-face communication, albeit a variant that has been subjected to many additional constraints (and especially constraints that involve the function of the communicative articulators).

Two facets of multimodal communication are of interest. The first involves examining multimodal communicative behaviour at a meta-level. For instance, given a characterization of a communicator (which includes the communicator's beliefs about the addressee and the communicative scenario, as well as some kind of desire to utter something), under sufficiently defined circumstances, can we characterize the multimodal utterances that might potentially be communicated? What measures should be used in order to characterize these multimodal utterances in terms of cost of articulation and potential to be misunderstood? Also, of particular interest to us, how do the potential multimodal utterances change as the properties of the communicator's articulators vary?

The second facet involves modeling the actual communicative behaviour of a particular communicator. This can be conceptualized as determining the best candidate multimodal utterance from among the potential multimodal utterances. What technique should be used for this selection? (We do not suggest that this is actually the process whereby a communicator's multimodal utterance is selected.) Such a mechanism for generating multimodal utterances is a generalization of natural language generation. We will not be concerned with this facet here.

The entities and actions described below have been implemented in Java, with the exception of the constraint satisfaction module, which has been implemented in a variant of Lisp.

3.1 Simulation Design

Face-to-face communication is certainly a collaborative process, and in particular, it requires the establishment of common ground (Clark and Wilkes-Gibbs, 1992). But, at the same time, the actual *articulation* of utterances is a task performed by an individual. If we want to examine the consequence of the *form* of the articulated utterance on the communicative exchange (e.g., in terms of communicator fatigue, potential for being misunderstood, burden on the addressee, or even whether communication breakdown occurs), then many of the dynamic aspects of communication must be held fixed. For this simulation, a sufficiently narrow communicative task is required so that the effect of this independent variable can be examined. The task of *definite reference communication* was selected for several reasons. First, the findings from a relatively constrained task (e.g., one in which the set of potential referents is relatively small) are scalable up to larger tasks. Second, the communication of definite reference is an important constituent sub-process in face-to-face conversation, and in particular, is especially fundamental to many AAC-system users. Third, it is a communicative act that is naturally multimodal because of the deictic reference involved. Last, it is a reasonable starting point for a simulation, as other sub-processes can be included later. Therefore, most discourse functions — such as turn-taking, back-channeling, signaling and repairing of misunderstanding — have been held fixed. Instead, the task facing the communicative agent is to perform the following *communication game*: first, select a referent from the environment, and then convey that information, by means of a multimodal utterance, to another communicative agent, which acts as a tester. An oracle monitors the exchange and gathers information about what multimodal utterance was performed by the communicative agent, and whether the definite referent was communicated successfully (this requires access to the communicative agent's private knowledge about the intended referent).

For this specific task, the implemented behaviour consists of *generating* semantic representations and then *"performing"* a corresponding multimodal utterance.

3.2 Characteristics of Computational Communicative Agents

In order to simulate the communicative behaviour of a range of communicators (e.g., those using any of a variety of AAC systems, or no AAC system at all), certain properties of the communicator's articulators must be parameterized. We emulate the behaviour of a communicator by means of a computational, communicative agent, which is defined as having:

- a set of modes $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$, with:
  - a **cost** function C, where $C : \mathcal{M} \longrightarrow [0, 1]$. A value $C(M_k) = 1$ indicates that the mode M_k is most costly, while values tending to 0 indicates that it is less costly.
  - a **unreliability** function R, where $R : \mathcal{M} \longrightarrow [0, 1]$. A value $R(M_k) = 1$ indicates that M_k is most unreliable, while values tending to 0 indicate lower unreliability (i.e., higher reliability).
- an **interference set** $I = \{I_1, I_2, \ldots, I_n\} \subset \mathcal{P}(\mathcal{M})$, which is used to represent the modes that conflict with one another.

  For each mode M_k, set $I_k \in I$ represents the modes which directly conflict with mode M_k and cannot be used simultaneously with it. (This set may be null.)

In order to simulate the communication of AAC-system users, the explicit definition of modes and the definition of the interference set are important parameters. In such communication, often an *aided* mode of communication is made available through the use of an AAC device. For instance, the aided mode of speech is made available through a speech synthesizer. Such aided modes have a high cost, however. In addition, to operate the AAC device, the user must look down (it is typically held on the lap or mounted on the frame of a wheelchair) and provide input actions in the form of keyboard or touch-screen presses. So, the use of the mode of gesture and the mode eye gaze are, in many cases, in conflict with the aided mode of speech, due to the style of interaction that is demanded by the AAC device.

One would expect that the correlation of these function values is negative. The less effort a mode requires, the more unreliable it is. However, this is not always the case. For instance, vocalization to an AAC system user requires a high effort level, yet is still unreliable.

3.3 Semantic Representation

In order to simulate the articulation of multimodal utterances, the pre-linguistic and pre-articulated form of an utterance must be represented. Although this representation captures semantic content, it is not complete, as the form of an utterance is also influenced by the communicator's perception of the communicative situation and the addressee. Furthermore, these percepts are not represented in the semantic representation, but rather come to bear when a particular semantic representation is uttered in a communicative act. As described previously, we are concerned with the task of definite reference communication, so the semantic domain is restricted to the potential referents (as defined to exist in the simulated environment).

The semantic representation of a referent X_j (where j represents an independent label for the referent) is characterized by:

- a set of **constituents** $X_j = \{c_1, c_2, \ldots, c_n\}$; and
- an ordering relation R_j on the set X_j (which is a non-reflexive, antisymmetric, but transitive relation).

The communication of each constituent c_i serves to provide information that disambiguates the intended referent from the set of potential referents. The ordering relation is a generalization of the ordering relation that is observed to hold for linguistic modes (e.g., one might say "the big red ball", and not "the red big ball"). Each of the constituents is articulatable by one or more modes, sequentially or simultaneously. The act of uttering all of the constituents conveys the intended definite referent. The omission of some of the constituents, or their articulation by particularly unreliable modes, results in a communicative act that conveys an ambiguous definite reference.

To capture these effects, each constituent is defined to have the following:

- an **effort** function E, where $E : X \times \mathcal{M} \longrightarrow (0, 1]$. A value $E(c_i, M_k) = 1$ indicates that to articulate c_i with respect to mode M_k requires maximum effort. $E(c_i, M_k) > 0$, since every articulates takes *some* amount of effort.

  In general, some types of semantic content are most easily expressed by speech (e.g., a description of an abstract property), while other types (e.g., the manner of motion) are most easily expressed gesturally. In this domain of definite reference, the differences are not so marked.

- an **uncertainty** function U, where $U : X \times \mathcal{M} \longrightarrow [0, 1]$. A value $U(c_i, M_k) = 1$ indicates that articulating constituent c_i with respect to mode M_k is extremely risky and almost certain to result in misunderstanding.

3.4 The Mechanism for Generating Semantic Representations

As described earlier, the context for the performance of communicative acts is a communication game in which the communicative agent must select a referent and then convey that information, through the performance of a multimodal utterance, to the simulated tester.

The process of generating semantic representations corresponds to the communicative agent "thinking" and "selecting" a referent to communicate. This process is approximated with a random choice mechanism operating over a set of pre-defined semantic representations.

3.5 Characteristics of Utterances

We have used a variety of representations for multimodal communicative acts, but, as shown in figure 1.1, timeline-based representations are most intuitive. The columns represent time-steps (in these abstract examples, the time units are arbitrary), and the rows represent the various modes. In these examples, we can imagine mode `m1` representing eye gaze, mode `m2` representing speech, and mode `m3` representing gesture. This representation is simply a matrix, say $\mathcal{A}$, with the rows and columns labelled with the modes and time-steps, respectively.

The matrix $\mathcal{A}$ is defined as follows:

$$A[i,j] = \begin{cases} \varnothing & \text{no constituent to be articulated using mode } i \text{ at time step } j, \\ k & \text{constituent } x_k \text{ to be articulated using mode } i \text{ at time step } j. \end{cases} \tag{1.1}$$

```
m1:        22      m1:                m1:111111
m2:1111 2222       m2: 1111 2222      m2: 1111 2222
m3:          33    m3:11111     33    m3: 11111    33
   123456789012       123456789012       123456789012
```

Figure 1.1: Three multimodal utterances represented in a timeline notation.

3.6 The Mechanism for Producing Utterances

The "performance" of a multimodal communicative act that realizes a particular semantic representation X_j (and takes into account other factors) can be conceptualized as the articulation of a multimodal utterance, where a multimodal utterance can be represented by a matrix $\mathcal{A}$ (as defined previously).

The mapping from semantic representation to utterance has been formalized as a constraint satisfaction problem and has been implemented in SCREAMER, a variant of Lisp that incorporates a non-deterministic choice operator and logic variables, thus serving as a substrate for constraint logic programming.

The variables of this constraint satisfaction problem represent, for each constituent of the semantic representation, any time-steps at which each of the defined modes is used for its articulation. In order for the domain for each of these variables to be finite, a simplifying assumption was made that for each communicative scenario, an utterance should not take more than $\mathcal{T}$ time-steps, where the value of $\mathcal{T}$ is an input to the CSP. The granularity of the time-steps can be arbitrarily small, although this has an impact on the tractability of the CSP.

There are several n-ary logical relations that must hold between these variables. (For brevity, only a summary of the constraints is given here, but the interested reader is directed to Baljko (2000) for a more thorough description.)

1. **Completeness**: Each $c_i \in X_j$ is expressed via *at least one* mode $M_k \in \mathcal{M}$;
2. **Conformity to Ordering Relation**: The sequence in which the c_i's are communicated obeys the ordering relation R_j; and
3. **Non-overlapping Mode Use**: A mode M_k can be used to articulate at most one constituent at a time.

A matrix $\mathcal{A}$ represents a valid or meaningful multimodal utterance if none of these criteria are violated. As one might expect, the problem is under-constrained. Therefore, as an intermediate step, the set of potential multimodal utterances for a given agent, in a given communicative scenario, is represented. As discussed in section 3.1, these representations are relevant for one facet of the study of multimodal communication — the meta-level in which one considers the characteristics of potential multimodal utterances.

A subsequent step in the performance of an utterance is the selection of a particular multimodal utterance from this set of candidates. This process is of interest in the other facet of the study of multimodal communication.

3.7 Characterizing Multimodal Utterances

For each multimodal utterance, we would like to determine values for its cost of articulation, its likelihood of being misunderstood and its degree of multimodality (e.g., the number of modes that are involved at each time-step). For each of these constructs, we have developed and implemented a number of different indicators, some of which are in table 1.1. We are currently experimenting with these indicators to determine their validity (i.e., the extent to which the indicator actually measures the qualitative property that it is purporting to determine).

4 Evaluation

In the previous sections, several parameters for both the communicative agent and the semantic representation have been abstracted. We want to separate the evaluation of the abstraction itself from the evaluation of the *particular* parameter values.

At this initial stage, the criterion for the abstraction is whether the inter-relationships between the measures described in the previous section agree with our expectations.

Although the particular values assigned to the parameters are arbitrary, we will demonstrate that the parameters themselves are meaningful, and, provided the values obey certain conditions, the overall emergent behaviour of the simulation is principled.

Property	Determined By
Cost of Articulation	$\alpha_1 = \frac{1}{\lvert\mathcal{X}\rvert}\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\sum_{M_j\in\mathcal{M}_{c_i}} C(M_j)\right]$ — the average cost to articulate a constituent (as determined by the modes' cost function). $\alpha_2 = \frac{1}{\lvert\mathcal{X}\rvert}\cdot\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\cdot\sum_{M_j\in\mathcal{M}_{c_i}} C(M_j)\cdot E(c_i, M_j)\right]$ — the average cost to articulate a constituent (as determined by the modes' cost function and compounded by the effort function of the constituent). $\alpha_3 = \frac{1}{\lvert\mathcal{X}\rvert}\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\sum_{M_j\in\mathcal{M}_{c_i}} C(M_j)\cdot E(c_i, M_{c_i})\cdot\text{simultaneity penalty for } c_i\right]$ — the average cost to articulate a constituent (as determined by the modes' cost function and compounded by the effort function of the constituent). If a constituent is articulated simultaneously by more than one mode, then the constituent's cost of articulation is adjusted to increase the average, as calculated by the modes' cost function and compounded by the effort function of the constituent.
Potential for Misunderstanding	$\beta_1 = \frac{1}{\lvert\mathcal{X}\rvert}\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\sum_{M_j\in\mathcal{M}_{c_i}} R(M_j)\right]$ — the average likelihood that a constituent will be misunderstood (as determined by the modes' unreliability function). $\beta_2 = \frac{1}{\lvert\mathcal{X}\rvert}\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\sum_{M_j\in\mathcal{M}_{c_i}} R(M_j)\cdot U(c_i, M_{c_i})\right]$ — the average likelihood that a constituent will be misunderstood (as determined by the modes' unreliability function and compounded by the uncertainty function of the constituent). $\beta_3 = \frac{1}{\lvert\mathcal{X}\rvert}\sum_{c_i\in\mathcal{X}}\left[\frac{1}{\lvert\mathcal{M}_{c_i}\rvert}\sum_{M_j\in\mathcal{M}_{c_i}} R(M_j)\cdot U(c_i, M_{c_i})\cdot\text{simultaneity reward for } c_i\right]$ — the average likelihood that a constituent will be misunderstood (as determined by the modes' unreliability function and compounded by the uncertainty function of the constituent). If a constituent is articulated simultaneously by more than one mode, then the constituent's likelihood of being misunderstood is adjusted to be lower than the average, as calculated in β_2.
Degree of Multimodality	$\mu_1 = \frac{1}{\lvert\mathcal{M}\rvert\cdot\lvert\mathcal{T}\rvert}\sum_{c_i\in\mathcal{X}}\sum_{M_j\in\mathcal{M}_{c_i}}(\text{number of timesteps mode } M_j \text{ is used for } c_i)$ — the proportion of mode use to maximum mode use (e.g., if all the modes had been used at all the time-steps).

Table 1.1: Various measures of multimodal utterances

4.1 Meta-properties of Multimodal Utterances

We ran a series of simulations to investigate the characteristics of the multimodal utterances produced.

In this series, the parameter values were assigned as follows:

- the function values of the unreliability and cost functions of the mode set $\mathcal{M}$ were inversely proportional;
- three modes of articulation were defined: `m1`: low effort but high unreliability, `m2`: moderate effort and moderate unreliability, and `m3`: high effort and low unreliability;
- the function values of the constituents' effort and uncertainty functions were inversely proportional.

In this series, the simulation environment was defined to contain eight potential referents, each with an equal number of constituents, and each characterised by the same effort and uncertainty functions.

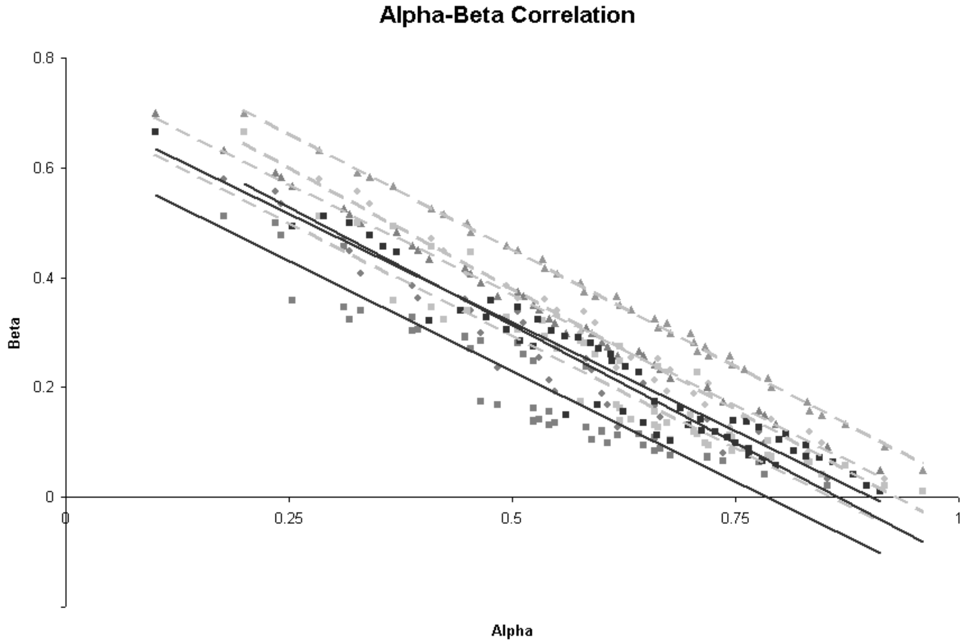

Figure 1.2: The cost of multimodal utterances (the α's) and the likelihood of being misunderstood (the β's) are negatively correlated.

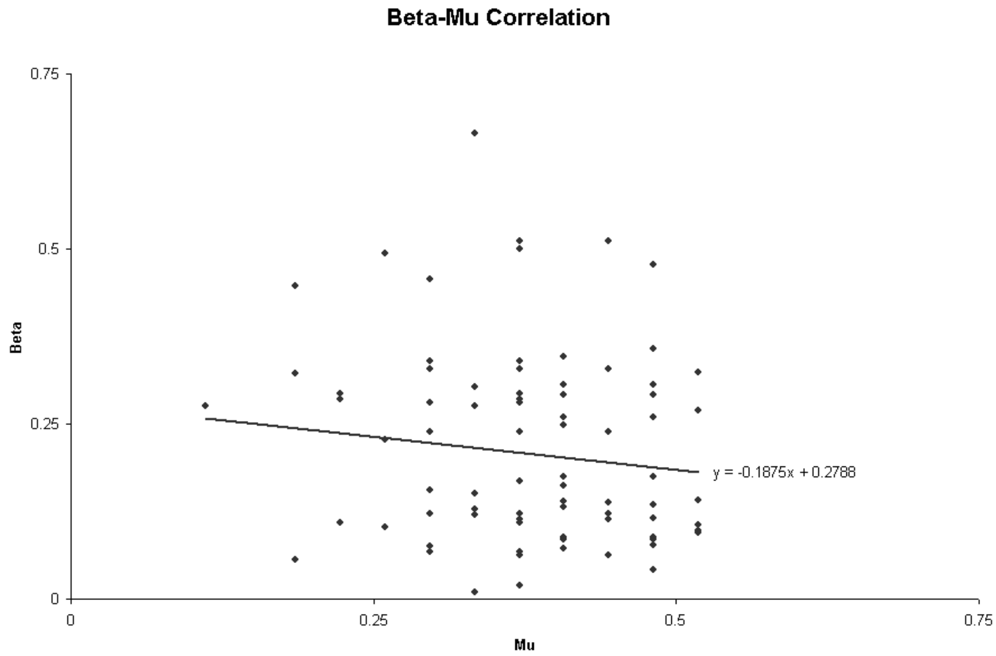

Figure 1.3: The degree of multimodality (μ) and the likelihood of being misunderstood (the β's) are negatively correlated.

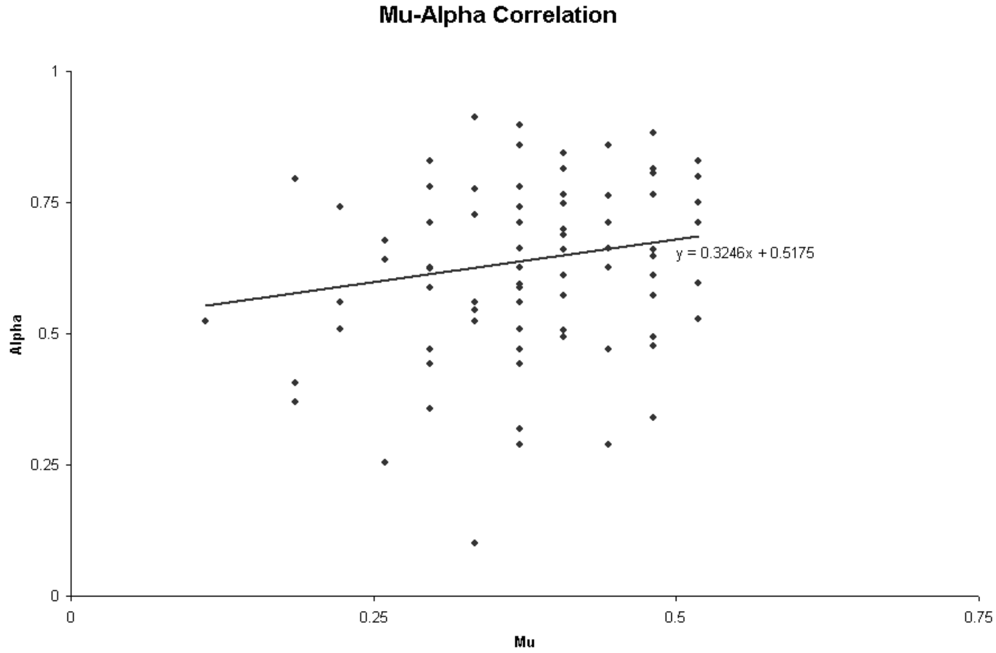

Figure 1.4: The cost of multimodal utterances (the α's) and the degree of multimodality (μ) are positively correlated.

Our analysis of the simulation output demonstrates that the following meta-properties can be found in the generated multimodal utterances:

- **The trade-off between the cost of articulation and the likelihood of being misunderstood.** As the cost of a multimodal utterance increases (measured by α's), one would expect that there to be a reduction in the likelihood of being misunderstood (measured by β's). This relationship is demonstrated in figure 1.2.

- **The trade-off between the degree of multimodality and the likelihood of being misunderstood.** The consequence of using multiple modes (where the degree of multimodality is measured by μ) should be the conveyance of additional information, which reduces the likelihood of being misunderstood (measured by β's). This relationship is demonstrated in figure 1.3.

- **The inter-relationship between degree of multimodality and the cost of articulation.** The higher the degree of multimodality (measured by μ), the more the articulators are used, which results in a higher articulation cost (measured by α). This relationship is demonstrated in figure 1.4.

In another series of simulations, we decided to vary the properties of the modes over time. At the initial time-step, all the modes were defined with extremely low costs and high unreliability values. The cost function values then were increased linearly over time, while a simultaneous linear decrease was applied to the values of the unreliability function. At each time-step, the potential multimodal utterances were evaluated with respect to cost of articulation and their likelihood of being misunderstood. In figure 1.5, the mean α_3 (`a3mean`) and β_3 (`b3mean`) values were shown at each time-step (as calculated over the range of all multimodal utterances that the communicative agent could potentially make at a particular time-step). Also shown are the upper and lower bounds for these measures (`a3max`, `b3max`, `a3min`, and `b3min`, respectively).

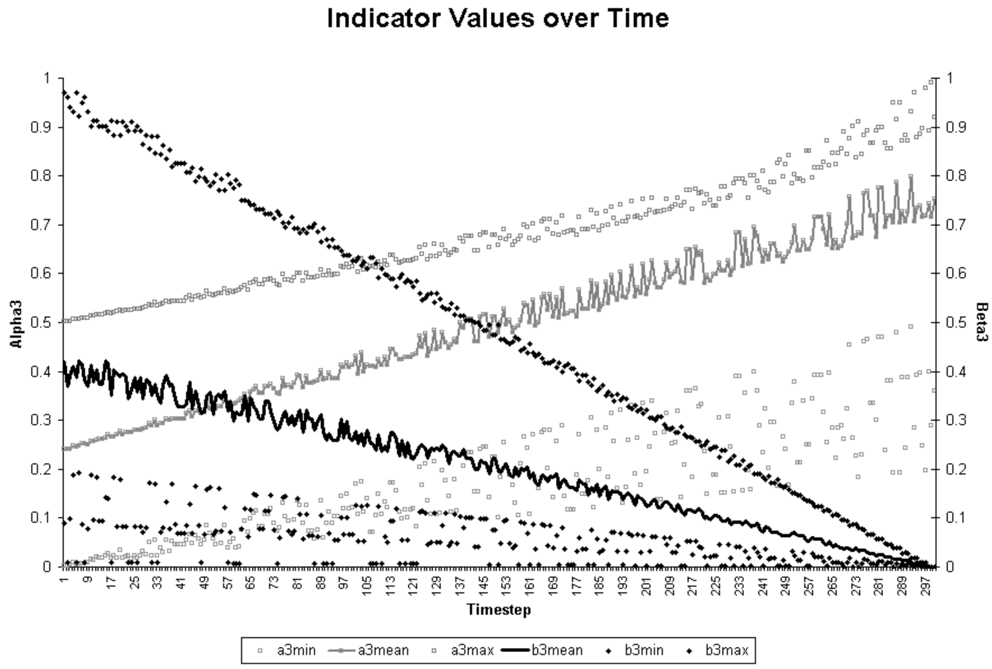

Figure 1.5: The cost of multimodal utterances (α_3) and the likelihood of being misunderstood (β_3) are negatively correlated, for an agent whose characteristics have been varied over time (e.g., the cost and unreliability functions for the set of modes were varied over time).

5 Empirical Validation

The representations of multimodal utterances that are generated by the simulation are comparable with empirical data gathered from digitized video that has been hand-coded by multiple judges. We have digitized multiple analog video-tape recordings of conversational dyads between AAC-system users and non-AAC-system users (this data was gathered by another researcher). We are coding and annotating, by hand on a frame-by-frame basis, the communicative behaviour of the interlocutors. We have adapted observational data analysis software, designed for ergonomic analysis in the human factors research community, for this purpose (Sanderson et al., 1994).

6 Conclusions and Future Work

To date, we have met our primary goal that the model be descriptive — that it account for the empirical data that we have gathered and the findings from the relevant research literature. As part of ongoing research, we are exploring whether the model can be used to predict the behaviour of communicators. The model has been embedded into two agents, which are operating within a simulated environment. We are currently exploring (and will report on) the effect of the manipulation of the communicator-dependent constraints on the communication strategy, the conditions under which communication breakdown occurs, and the strategies the agents use.

Acknowledgments I am grateful to Graeme Hirst for his comments on earlier drafts of the work. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Bibliography

Arens, Y. and Hovy, E.: 1995, The design of a model-based multimedia interaction manager, *Artificial Intelligence Review* 9(2–3), 167–188

Austin, J. L.: 1962, *How to Do Things with Words*, Harvard University Press

Baljko, M.: 2000, MMSIM*: Simulated Multimodal Communication*, Technical Report CSRG-411, University of Toronto, Toronto, Canada, Available at ftp://ftp.cs.toronto.edu/pub/reports/csri/411/

Baljko, M.: forthcoming, What is a multimodal speech act?, in *submitted*

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., and Achorn, B.: 1994, Modeling the interaction between speech and gesture, in *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology, Atlanta, USA

Clark, H. H. and Wilkes-Gibbs, D.: 1992, Referring as a collaborative process, in *Arenas of Language Use*, Chapt. 4, pp 107–143, The University of Chicago Press

Johnston, M.: 1998, Unification-based multimodal parsing, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, Vol. 1, pp 624–630, Montreal, Canada

McGee, D. R., Cohen, P. R., and Oviatt, S.: 1998, Confirmation in multimodal systems, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, Vol. 2, pp 823–829, Montreal, Canada

Newell, A. F., Arnott, J. L., Cairns, A. Y., Ricketts, I. W., and Gregor, P.: 1995, Intelligent systems for speech and language impaired people: A portfolio of research, in A. D. N. Edwards (ed.), *Extra-Ordinary Human-Computer Interaction: Interfaces for Users with Disabilities*, Chapt. 5, pp 83–101, Cambridge University Press

Oviatt, S.: 1999, Mutual disambiguation of recognition errors in a multimodal architecture, in *Proceedings of CHI 99, the 1999 Conference on Human Factors in Computing Systems*, pp 576–583

Sanderson, P. M., Scott, J. J. P., Johnston, T., Mainzer, J., Watanabe, L. M., and James, J. M.: 1994, MacSHAPA and the enterprise of Exploratory Sequential Data Analysis (ESDA), *International Journal of Human-Computer Studies* 41, 633–668

Searle, J. R.: 1979, *Expression and Meaning*, Cambridge University Press