MODELING THE INTERACTION BETWEEN SPEECH AND GESTURE

Justine Cassell *Mark Steedman* *Norm Badler* *Catherine Pelachaud*
Matthew Stone *Brett Douville* *Scott Prevost* *Brett Achorn*

Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389
Telephone: (215) 573-2821
Fax:(215) 898-0587
justine@central.cis.upenn.edu*

February 14, 1994

Abstract

Until now theories of the gesture-speech relationship have been difficult to evaluate because of their descriptive basis. In this paper we provide a tool for investigating the relationship between speech and gesture: a system that generates speech, intonation, and gesture using two copies of an identical program that have different knowledge of the world and must cooperate to accomplish a goal. The output of the dialogue generation is fed into a three-dimensional interactive animated model – two graphic figures on a computer screen who gesture according to the rules given to the system. The advantage of computer modeling in this domain is that it forces us to come up with predictive theories of the gesture-speech relationship. A felicitous outcome is a working system to realize autonomous animated conversational agents, for virtual reality and other purposes.

Areas: discourse, gesture, animation Presentation: talk

*This research is partially supported by NSF VPW GER-9350179; NSF IRI91-17110; NSF graduate fellowships; ARO Grant DAAL03-89-C-0031 including participation by the U.S. Army Research Laboratory (Aberdeen); U.S. Air Force DEPTH contract through Hughes Missile Systems F33615-91-C-000; DMSO through the University of Iowa; National Defense Science and Engineering Graduate Fellowships; Naval Training Systems Center N61339-93-M-0843; Sandia Labs AG-6076; NASA KSC NAG10-0122; MOCO, Inc.; and NSF Instrumentation and Laboratory Improvement Program Grant USE-9152503.

Modeling the Interaction between Speech and Gesture

Justine Cassell *Mark Steedman* *Norm Badler* *Catherine Pelachaud*
Matthew Stone *Brett Douville* *Scott Prevost* *Brett Achorn*

Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389

1 Introduction

In this paper we provide evidence that gesture and speech are different communicative manifestations of one single mental representation by attempting to model the interaction between them. Research on the relationship between gesture and speech has been difficult to evaluate because of its descriptive basis. One way to move from descriptive to predictive theories is via formal models, which point up gaps in knowledge and fuzziness in theoretical explanations. We have provided such a model, a dialogue generation program that drives two animated human figures, thus simulating the generation and carrying out of conversational interaction. In the remainder of the introduction we describe research on the relationship between gesture and speech that underlies our attempt to simulate the behavior. We follow that with a discussion of intonation and information structure, and give a set of rules for gesture generation with respect to those two linguistic variables. We then describe the various modules of the simulation: the dialogue generation program, speech and intonation synthesis, gesture integration, and animation interface.

Four basic types of gestures occur only during speaking ([13]); these four types of speech-associated gesture have been the focus of the majority of research on the cognitive basis of the gesture-speech relationship, including our own. *Iconics* represent some feature of the accompanying speech, such as sketching a small rectangular space with one's two hands while saying "Did you bring your CHECKBOOK?". *Metaphorics* represent an abstract feature concurrently spoken about, such as forming a jaw-like shape with one hand, and pulling it towards one's body while saying "You must WITHDRAW money.". *Deictics* indicate a point in space. They accompany reference to persons, places and other spatializeable discourse entities. An example is pointing to the ground while saying "Do you have an account at Mellon or at THIS bank?". Finally, *Beats* are

small formless waves of the hand that occur with heavily emphasized words, occasions of turning over the floor to another speaker, and other kinds of special linguistic work. An example is waving one's left hand briefly up and down along with the stressed words in the phrase "Go AHEAD."

Evidence from many sources suggests a close relationship between speech and gesture. At the prosodic level, [10] found that the stroke phase (most effortful part) of these gestures tends to co-occur with or just before the phonologically most prominent syllable of the accompanying speech. At a cognitive level, [4] established that listeners rely on information conveyed in gesture as they try to comprehend a story; [1] showed that children may express in gesture information that they cannot yet express in speech. Other evidence comes from the sheer frequency of gestures during speech. About three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another [13], and perhaps surprisingly, although the proportion of gesture types may change, all of these gestures, and spontaneous gesturing in general, are found in discourses by speakers of most languages.

In this paper, however, our primary concern is with the semantic and pragmatic relationship between the two media. Gesture and speech do not always manifest the same information about an idea, but what they convey is always complementary. That is, gesture may depict the way in which an action was carried out when this aspect of meaning is not depicted in speech. It has been suggested ([11]) that those concepts difficult to express in language may be conveyed by gesture. Thus simultaneity of two events, or the respective locations of two objects may be expressed by the position of the two hands. In this sense, the gesture-speech relationship resembles the interaction of words and graphics in the generation of multimodal text ([6],[21]). In storytelling, narrative structure may be indexed by differential use of gesture: iconic gestures tend to occur with plot-advancing description of the action, deictic gestures with the introduction of new characters, and beat gestures at the boundaries of episodes ([3]).

We propose to use the level of *information structure* to capture regularities such as these. The information structure of an utterance defines its relation to other utterances in a discourse and to propositions in the relevant knowledge pool. Although a sentence like "George withdrew fifty dollars" has a clear semantic interpretation, the semantics does not indicate how the proposition

relates to other propositions in the discourse. For example, the sentence might be an equally appropriate response to the questions "Who withdrew fifty dollars," "What did George withdraw," "What did George do," or even "What happened." Which question is asked determines which items in the response are most important or salient, which in turn determines how the phrase is uttered. These types of salience distinctions are encoded in the information structure representation of an utterance.

Following Halliday and others ([9], [8]), we use the terms *theme* and *rheme* to denote two distinct information structural attributes of an utterance.[1] The theme roughly corresponds to what the utterance is about. The rheme corresponds to what the speaker has to contribute concerning the theme. Depending on the discourse context, a given utterance may be divided on semantic and pragmatic grounds into thematic and rhematic constituents in a variety of ways. For example, given the utterance "George withdrew fifty dollars," we might consider the theme to be 'How much money George withdrew' and the rheme to be 'fifty dollars.'

Within information structural constituents, we define the semantic interpretations of certain items as being either *focused* or *background*. Items may be focused for a variety of reasons, including emphasizing their newness in the discourse or making contrastive distinctions among salient discourse entities. For example, in a theme concerning 'How much money George withdrew' we may say that 'George' may be the focus because it stands in contrast to some other salient discourse entity, say 'Gilbert'. We also mark the representation of entities in information structure with their status in the discourse. Entities are considered either new to discourse and hearer (indefinites), new to discourse but not to hearer (definites on first mention), or old (all others) ([18]).

Distinct intonational tunes have been shown to be associated with the thematic and rhematic parts of an utterance for certain classes of dialogue ([16],[17],[19]). In particular, we note that the standard rise-fall intonation generally occurs with the rhematic part of many types of utterances. The rise-fall intonation is realized as a pitch peak on the primary-stress syllable of the focused word, followed by an immediate fall to a lower pitch which is then sustained for the duration of the

[1] Although note that we drop Halliday's assumption that themes occur only in sentence–initial position. Functionally similar distinctions in this context are *topic/comment*, *given/new*, and the scale of *communicative dynamism*.

phrase. The rhematic part of yes/no interrogatives is often accompanied by a fall-rise intonation, realized as a low pitch target on the primary-stress syllable of the focused word, followed by an immediate rise to a sustained higher pitch. Thematic elements of an utterance are often marked by a rise-fall-rise intonation, realized by a rise to a high pitch target on the primary-stress syllable, followed by an immediate fall to a lower pitch with another pitch rise occurring at the end of the phrase. The following examples illustrate the coupling of tunes with themes and rhemes.

(1) Q: I know who withdrew three dollars, but who withdrew fifty dollars?
A: (GEORGE)$_{\text{RHEME}}$ (withdrew FIFTY dollars)$_{\text{THEME}}$

(2) Q: I know how many dollars Gilbert withdrew but how many did George withdraw?
A: (GEORGE withdrew)$_{\text{THEME}}$ (FIFTY dollars)$_{\text{RHEME}}$

The rules we propose to predict the location and types of gestures are these:

- Non-beat gestures accompany verb phrases in the rheme, and hearer new references, as follows: words with literally spatial content get iconic gestures; those with metaphorically spatial content get appropriate metaphorics; words with spatializable content get deictics.
- Beat gestures are generated for verb phrases in the rheme and for hearer new references when the semantic content cannot be represented spatially.
- Beats accompany discourse new definite references.

Generated gestures associated with a word are aligned with the stressed syllable of that word. This is a straightforward task for beat gestures which are simply waves of the hand. In contrast, the other gestures have a preparation phase which occurs before the stroke; accordingly, gestures with a preparation phase must start at the beginning of the intonational phrase in which the associated word occurs to ensure that the stroke can occur on that word.

2 Implementation

The objective of this project is to provide a testbed where predictive accounts of gesture, such as the rules given above, can be formalized and evaluated. This goal places certain demands on the generation of content for the "computational stage" ([13]) shared by speech and gesture. In particular, the generation process must provide precise and explicit representations of the concepts, such as information structure, to which the theory of gesture refers.

The solution adopted here is to simulate the world and discourse actions of an agent interacting with another agent in the service of accomplishing a goal in a simple environment. For the present

implementation, the ‘bank domain’ was chosen because in it there are two agents who must interact linguistically to accomplish a goal. The complexity of the domain and the steps followed by the agents are analogous to those of Power ([15]), but here the model is enriched with explicit representations of the structure of discourse and the relationship of the structure to the agents’ domain plans. The following two sections of the paper describe the dialogue generation system in this light.

The selection of content for the dialogue in our system is performed by two cascaded planners. The first is the domain planner, which manages the plans governing the concrete actions which the agents will execute; the second is the discourse planner, which manages the communicative actions the agents must take in order to agree on a domain plan and in order to remain synchronized while executing a domain plan.

The input to the domain planner is a database of facts describing the way the world works, the goals of the agents, and the beliefs of the agents about the world, including the beliefs of the agents about each other. The domain planner executes by decomposing an agent’s current goals into a series of more specific goals according to the hierarchical relationship between actions specified in the agent’s beliefs about the world. An agent’s goals may be of one of two forms: to obtain some piece of information, or to ensure that some state holds in the world; questions can be used to achieve either kind of goal, but planning decompositions are only appropriate for the second kind. Once decomposition resolves a plan into a sequence of actions to performed, the domain planner causes the agents to execute those actions in sequence. As these goal expansions and action executions take place, the domain planner also dictates discourse goals that agents must adopt in order to maintain and exploit cooperation with their conversational partner.

The domain planner transmits its instructions to take communicative actions to the discourse planner by suspending operation when such instructions are generated and relinquishing control to the discourse planner. Several stages of processing and conversational interaction may occur before these discourse goals are achieved. The discourse planner must identify how the goal submitted by the domain planner relates to other discourse goals that may still be in progress. Then content for a particular utterance is selected on the basis of how the discourse goal is decomposed into

sequences of actions that might achieve it.

The domain planner and the discourse planner offer a number of explicit representational structures which could serve as input in formulating rules of gesture and intonation. At any point, of course, each agent has a representation of the domain plan that is being executed, and of the constituents of discourse that go into discussion of the plan. Explicit links between these two structures indicate what part of the plan each discourse segment concerns; these links ensure that conversation is coordinated and understood. These three kinds of information form the basis for two additional levels of representation, which are maintained solely for their possible relevance to linguistic processes. First, a model of attention (the attentional state) indicates which entities are known to the participants, which entities have been referred to, and how salient those entities are. The attentional state for some utterance in the discourse consists of a list that contains, for each discourse segment which dominates that utterance, the sets of entities mentioned in that segment. These sets are ordered so that the entities referred to in larger segments are less salient than the entities referred to in segments they dominate. Second, a record of the purposes generated by the planner which initiated discourse actions is kept. Of course, it may happen that only the agent who initiated an action knows this purpose exactly. Accordingly, both parties also separately record the most specific purpose for a segment for which evidence has been given. This architecture of intentional structure, attentional state, and discourse purposes, and the relationship between them was first proposed by [7]; the implementation of these notions here follows their suggestions as closely as possible. We use these representations to reconstruct the information structure of the dialogue as follows.

- Material is classified as thematic if it occurs in some part of the speaker's discourse purpose in the current constituent or its ancestors for which evidence has been given.
- Material is classified as rhematic if it occurs only in that part of the speaker's discourse purpose in the current segment or its ancestors for which evidence has not been provided.
- Information not meeting either of these criteria constitutes linguistic formulae, which are irrelevant to the speaker's purpose, and are also labelled as thematic.

Focus is assigned to references according to the theory of contrast in [16], while the discourse status of entities is determined from the agents' knowledge of each other and from the attentional

state. Finally. the semantic class of constituents is retrieved from a dictionary associating semantic representations with possible gestures that might represent them.[2]

These structures permit application of the rules for generation of gestures and intonation given above. A variant of Prevost and Steedman's algorithm is used to do this, thus generating English text annotated with intonational cues and gestural instructions from information structures. These intonational and gesture features are attached to words in the dialogue and may alternatively be interpreted as occurring at the start of the associated word, on the stressed syllable of the word, or at the end of the word, depending on the feature. In order for the gestures to appear at the proper times in the animation, the two streams must be synchronized with the synthesized speech.

The intonation stream provides an abstract representation which is automatically translated to a form suitable for input to the speech synthesis component. We currently use the AT&T Bell Laboratories TTS synthesizer to produce the actual speech wave and phoneme timings ([12]).

After transforming the utterances into proper input for the synthesizer and generating the speech wave and phoneme timings, the durational outputs from the synthesis are merged by rule with the abstract intonational and gestural notations. This detailed timing information (to the centisecond) allows synchronization of the gestural animations with the speech, as described below.

3 Gesture Integration and Animation

In the research presented here the interaction between speech and gesture is modeled in such a form that it can drive an animation system[3]. The input to the animation system should not specify every small movement of the hands because it is determined by semantics rather than physiology, and it should take into account temporal deformations of gestures due to the demands of synchronizing gestures with speech and with one another.

Gesture production is carried out by a group of Parallel Transition Networks (PaT-Nets), finite state machines several of which can be run in tandem ([2]). PaT-Nets govern three processes, two of which concern the direct production of gesture through the animation system. The first,

[2]This solution is provisional: a richer semantics would include the features relevant for gesture generation, so that the form of gestures could be generated algorithmically from the semantics. Note also, however, that following Kendon we are led to believe that gestures may be more standardized than previously thought[11].

[3]Another model currently in progress generates gaze and head movements, and synchronizes gestures with these facial parameters as well as with movements of the lips ([14],[5])

parse-net, is a control network which parses the output of the speech synthesis module described above. This finite state machine parses phoneme representations one utterance at a time; in the current domain, this means also that one speaker turn is parsed at a time.

Upon the signalling of a particular gesture, parse-net will instantiate one of two additional PaT-Nets; if the gesture is a beat, the finite state machine representing beats ("beat-net") will be called, and if a deictic, iconic, or metaphoric, the network representing these types of gestures ("gest-net") will be called. This separation is motivated by the “rhythm hypothesis” ([20]) which posits that beats arise from the underlying rhythmical pulse of speaking, while other gestures arise from meaning representations. In addition, beats are often found superimposed over the other types of gestures, and such a separation facilitates implementation of superposition. Finally, since one of the goals of the model is to reflect differences in behavior among gesture types, this system provides for control of freedom versus boundedness in gestures (e.g. an iconic gesture or emblem is tightly constrained to a particular standard of well-formedness, while beats display free movement); free gestures may most easily be generated by a separate PaT-Net whose parameters include this feature.

Gesture and beat finite state machines are built as necessary by the parser, so that the gestures can be represented as they arise. The newly created instances of the gesture and beat pat-nets do not exit immediately upon creating their respective gestures; rather, they pause and await further commands from the calling network, in this case, parse-net. This is to allow for the phenomenon of gesture coarticulation, in which two gestures may occur in an utterance without intermediary relaxation, i.e. without dropping the hands or, in some cases, without relaxing handshape. Once the end of the current utterance is reached, the parser adds another level of control: it forces exit without relaxation of all gestures except the gesture at the top of the stack; this final gesture is followed by a relaxation of the arms, hands, and wrists.

The animation itself is carried out by *Jack*™ [†], a program for controlling articulated objects, especially human figures. The figures have joints and behaviors designed to generate realistic motion. Additional modules can be added to deal with new domains, such as gesture.

The PaT-Net system issues gesture requests to the animation system, telling the figure to either

[†] *Jack* is a registered trademark of the University of Pennsylvania.

*********Figure 1: 4 frames showing 4 different gestures**********

Figure 1: Examples of rule-based gesture generation

rest, make a beat motion, or make a gesture involving the hand, wrist, and/or arm. Four motion modules have been added to the Jack system: hand motion, wrist motion, arm motion, and beat motion, each which may be specified separately for each arm. The animation system isolates the higher level PaT-Net system from the details of the human figure geometry, biomechanical modeling, and joint control functions.

The hand motions can be specified in terms of an expandable library of hand shapes, and the current system is based on the American Sign Language alphabet. An additional parameter controls the laxness of the handshape. The animation system system moves the fingers from one position to another, attempting to get as close to the goal positions as possible within the contraints of the time alloted and the velocity limits of the finger joints. The result is that as the speed of the gesture increases, the gestures will 'coarticulate' in a realistic manner.

The wrist position goals are specified in terms of the hand direction relative to the figure (e.g. fingers forward and palm up). The animation system automatically limits the wrist to a realistic range of motion. Beat motions are a specialized form of wrist motion. Rather than having the goal specified, the goal is automatically generated based on the current position of the wrist. The animation system selects the most comfortable way for the figure to gesture in that situation and moves the wrist accordingly. The arm motions are specified in a manner similar to that of the wrists, except that relative spatial positions (e.g. near to the body, far left, and chest high) are given instead of orientations.

4 Example of Output

In Figure I, we see examples of how gestures are generated from discourse content.

1. "Shall we make a PLAN?"
    - In the first frame, a metaphoric gesture (the common *conduit* gesture, representing the plan as an entity that can be presented to the listener) is generated because of the first mention (new to hearer) of the abstract notion 'make a plan'.

2. "I suggest that you WRITE a check"
    - In the second frame, an iconic gesture (representing writing on a piece of paper) is generated from the first mention of the concrete action of 'writing a check'.
3. "Your account contains THREE dollars"
    - In the third frame, an iconic gesture (the *emblematic* gesture understood to mean 'three') is generated from the first mention (new to hearer) of the entity 'three dollars'.
4. "Three dollars is LESS than fifty dollars"
    - In the fourth frame, a beat gesture (a movement of the hand up and down) is generated from the first mention of the notion 'less than', which cannot be represented spatially.

5 Conclusion

Most research on gesture has been descriptive and distributional. With the evidence available, it is time to attempt predictive theories of gesture use. The research on gesture-speech interaction described above was sufficient to allow us to specify rules and write algorithms that drive a animated model of verbal and non-verbal behaviors in conversational interaction. Formal models such as ours point up gaps in knowledge, and fuzziness in theoretical explanations. In the current case, we discovered that while we could quite successfully specify when gestures might be expected in a discourse, and what the temporal relationship between those gesture and speech would be, we lacked the knowledge to *distribute* communicative load among gesture, the semantics of speech, and intonation. That is, the discourse model generated turn-taking phrases such as "go ahead" and also generated beat gestures to accompany those phrases. In natural human interaction, it is more likely that either gesture or speech take on such a linguistic function, but not the two systems simultaneously. We are therefore led to return to the parallel established above with automatically generated coordinated multimedia presentations as a way of, in the future, improving the accuracy of our model. In the meantime, we have demonstrated that gesture and speech can be generated from one single underlying representation, and that they can be treated as an integrated conceptual system. And we have, in the process, produced autonomous animated conversational agents.

References

[1] M. W. Alibali and S. Goldin-Meadow. Modeling Learning using Evidence from Speech and Gesture. *Proceedings of the Annual Conference of the Cognitive Science Society*, 1993.

[2] Welton M. Becket. The *jack lisp api*. Technical Report MS-CIS-94-01/Graphics Lab 59, University of Pennsylvania, 1994.

[3] J. Cassell and D. McNeill. Non-verbal imagery and the poetics of prose. *Poetics Today*, 12(3):375–404, 1991.

[4] J. Cassell, D. McNeill, and K.-E. McCullough. Kids, don't try this at home: Experimental mismatches of speech and gesture. Presented at the International Communication Association annual meeting, 1993.

[5] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, C. Seah, and M. Stone. Animated conversation: Rule based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *SIGGRAPH'94*, 1994. (submitted).

[6] S. Feiner and K. McKeown. Automating the generation of coordinated multimedia explanations. *IEEE Computer*, 24(10), 1991.

[7] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

[8] E. Hajičová and P. Sgall. Topic and focus of a sentence and the patterning of a text. In János Petofi, editor, *Text and Discourse Constitution*. De Gruyter, Berlin, 1988.

[9] M. Halliday. *Intonation and Grammar in British English*. Mouton, The Hague, 1967.

[10] A. Kendon. Movement coordination in social interaction: some examples described. In Weitz, editor, *Nonverbal Communication*. Oxford University Press, 1974.

[11] A. Kendon. Do gestures communicate: A review. *Research on Language and Social Interaction*, 1994.

[12] M. Liberman and A. L. Buchsbaum. Structure and usage of current Bell Labs text to speech programs. Technical Memorandum TM 11225-850731-11, AT&T Bell Laboratories, 1985.

[13] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago, 1992.

[14] C. Pelachaud, N.I. Badler, and M. Steedman. Linguistic issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, 1991.

[15] R. Power. The organisation of purposeful dialogues. *Linguistics*, 1977.

[16] S. Prevost and M. Steedman. Generating contextually appropriate intonation. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pages 332–340, Utrecht, 1993.

[17] S. Prevost and M. Steedman. Using context to specify intonation in speech synthesis. In *Proceedings of the 3rd European Conference of Speech Communication and Technology (EuroSpeech)*, pages 2103–2106, Berlin, 1993.

[18] E. F. Prince. The ZPG letter: Subjects, definiteness and information status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins B.V., 1992.

[19] M. Steedman. Structure and intonation. *Language*, pages 260–296, 1991.

[20] K. Tuite. The production of gesture. *Semiotica*, 93(1/2), 1993.

[21] W. Wahlster, E. André, W. Graf, and T. Rist. Designing illustrated texts. In *Proceedings of the 5th EACL*, pages 8–14, 1991.