7-5-1996

Synthesizing Cooperative Conversation

Catherine Pelachaud
University of Pennsylvania

Justine Cassell
University of Pennsylvania

Norman I. Badler
University of Pennsylvania, badler@seas.upenn.edu

Mark Steedman
University of Pennsylvania, steedman@seas.upenn.edu

Scott Prevost
University of Pennsylvania

See next page for additional authors

Synthesizing Cooperative Conversation

Abstract
We describe an implemented system which automatically generates and animates conversations between multiple human-like agents with appropriate and synchronized speech, intonation, facial expressions, and hand gestures. Conversations are created by a dialogue planner that produces the text as well as the intonation of the utterances. The speaker/listener relationship, the text, and the intonation in turn drive facial expressions, lip motions, eye gaze, head motion, and arm gesture generators.

Disciplines
Computer Sciences | Engineering | Graphics and Human Computer Interfaces

Author(s)
Catherine Pelachaud, Justine Cassell, Norman I. Badler, Mark Steedman, Scott Prevost, and Matthew Stone

Synthesizing Cooperative Conversation

Catherine Pelachaud* Justine Cassell† Norman Badler Mark Steedman
Scott Prevost Matthew Stone
University of Pennsylvania ‡

July 5, 1996

Abstract

We describe an implemented system which *automatically* generates and animates conversations between multiple human-like agents with appropriate and synchronized speech, intonation, facial expressions, and hand gestures. Conversations are created by a dialogue planner that produces the text as well as the intonation of the utterances. The speaker/listener relationship, the text, and the intonation in turn drive facial expressions, lip motions, eye gaze, head motion, and arm gesture generators.

1 Introduction

Conversation is an interaction between agents, who cooperate to achieve mutual goal using spoken language (words and contextually appropriate intonation marking topic and focus), facial movements (lip shapes, emotions, gaze direction, head motion), and hand gestures (handshapes, points, beats, and motions representing the topic of accompanying speech). Without being able to deploy all of these verbal and non-verbal behaviors, a virtual agent cannot be realistic, believable. To limit the problems (such as voice and face recognition, and conversational inference) that arise from the involvement of a real human conversant we have developed a dialogue generation system in which two copies of an identical program, differing only in their specific knowledge of the world, must cooperate to accomplish a goal. Both agents of the conversation collaborate via the dialogue to construct a simple plan of action. They interact with each other to propose goals, exchange information, and ask questions.

The work presented builds on a considerable body of research on the relation between gesture, facial expression, discourse meaning and intonation. Most of this research has been purely descriptive lacking any formal theory relating form to discourse meaning. We present a fragment of a theory which allows us to control and verify the communicative models. The development of such a formal theory is as central a goal of the research as is its exploitation in animation. We have used it to develop a high-level programming language for 3D animation, which we view as a tool to investigate gestural and facial behavior together with spoken intonation, within the context of a dialog.

This language allows users to *automatically animate conversations between multiple human-like agents with appropriate and synchronized speech, intonation, facial expressions, and hand gestures*. In people, speech, facial expressions, and gestures are all the reflection of a single system of meaning. While an expert animator may realize this unconsciously in the "look" of a properly animated character, a program to automatically generate motions must embed this knowledge in a system of rules. A fuller report of some aspects of this work appears in [9] and [10].

1.1 Literature on Facial Animation

Various systems have been proposed for facial animation. Since animating faces manually is very tedious and requires the skill of a talented animator, parameterized, rule-based or analysis-based approaches have been studied.

*Universitá di Roma "La Sapienza"
†M.I.T., Media Lab
‡The authors would like to thank Brett Achorn, Tripp Becket and Brett Douville

The set of parameters from Parke's model [49] distinguished conformation parameters from expression parameters. This set was then extended by adding speech parameters to include lip synchronization [30, 51, 48, 12]. Cohen et al. [12] uses overlapping dominance functions to consider coarticulation phenomenon.

Rule-based systems are based on a set of rules to drive automatically the animation. Multi level structures offer a higher level animation language [35, 50]. At the lower level the deformation controller simulates muscle actions by moving some control points. At the higher level the expression controller defines facial expression and lip shape for sentences.

Greater realism at the expense of synthetic control comes from analysis-based techniques which extract information from live-animation. The computed movement information is interpreted as muscle contractions and is given as input to the animation system [21, 66, 40].

Takeuchi [64] propose a categorization of facial expressions depending on their communicative meaning, and implement this framework in a user-interface where a 3D synthetic actor recognizes the words pronounced by a user and generates a response with the appropriate facial displays.

1.2 Literature on Gesture Animation

A "Key-framing" technique is commonly used to create arm and hand motions. Rijpkema and Girard [60] created handshapes automatically based on the object being gripped. The Thalmanns [23, 44] improved on the hand model to include much better skin models and deformations of the finger tips and the gripped object. Lee and Kunii [38] built a system that includes handshapes and simple pre-stored facial expressions for American Sign Language (ASL) synthesis. Dynamics of arm gestures in ASL have been studied by Loomis et al [42]. Chen et al [11] constructed a virtual human that can shake hands with an interactive participant. Lee et al [39] automatically generate lifting gestures by considering strength and comfort measures. Moravec and Calvert [8] constructed a system that portrays the gestural interaction between two agents as they pass and greet one another. Behavioral parameters were set by personality attribute "sliders" though the interaction sequence was itself pre-determined and limited to just one type of non-verbal encounter.

1.3 Literature on Dialog and Intonation Generation

Generation of natural language is an active area of current research, comprising several independent subproblems. In a high-level phase, generation systems typically identify, organize and sequence material to be presented to determine the overall structure of a contribution to discourse (e.g., [34, 47]). An intermediate process then structures ideas into sentence-sized units and determines words and referring expressions to use (e.g., [15, 46]). A final phase realizes the resulting structures as sentences using linguistic knowledge (e.g. [62]). A good survey of how these tasks are performed in a several recent systems can be found in [59].

A few other researchers have attempted to automatically generate both sides of a dialogue [54, 31, 6, 24, 68] and to generate communication in the form of text and illustration presented simultaneously [22, 67]. Generating speech and gesture for conversations between two animated agents requires a synthesis of techniques in both areas.

A number of researchers have investigated the problem of automatically generating intonational contours in natural language generation systems. Early work by Terken [65] is concerned with determining relative levels of givenness for discourse entities and applying pitch accents accordingly. The Direction Assistance program, designed by Davis and Hirschberg [16] determines a route and provides spoken directions for traveling between two points on a map, assigning pitch accent in the synthesized speech based on semantic notions of givenness. Work by Houghton, Isard and Pearson [32, 33] undertakes the similar task of assigning intonation to computer-generated, goal-directed dialogues. More recently, Zacharski et al. [69] have proposed a system for generating utterances and appropriate intonation in map-task dialogues.

1.4 Example

In this section of the paper we present a fragment of dialogue (the complete dialogue has been synthesized and animated), in which intonation, gesture, head and lip movements, and their inter-synchronization were automatically generated. This example will serve to demonstrate the phenomena described here, and in subsequent sections we will return to each phenomenon to explain how rule-generation and synchronization are carried out.

In the following dialogue, imagine that Gilbert is a bank teller, and George has asked Gilbert for help in obtaining $50. The dialogue is unnaturally repetitive and explicit in its goals because the dialogue generation program that produced it has none of the conversational inferences that allow human conversationalists to follow leaps of reasoning. Therefore, the two agents have to specify in advance each of the goals they are working towards and steps they are following (see Section 2.1).

Gilbert: Do you have a blank check?
George: Yes, I have a blank check.
Gilbert: Do you have an account for the check?
George: Yes, I have an account for the check.
Gilbert: Does the account contain at least fifty dollars?
George: Yes, the account contains eighty dollars.
Gilbert: Get the check made out to you for fifty dollars and then I can withdraw fifty dollars for you.
George: All right, let's get the check made out to me for fifty dollars.

When Gilbert asks a question, his voice rises. When George replies to a question, his voice falls. When Gilbert asks George whether he has a blank check, he stresses the word "check". When he asks George whether he has an account for the check, he stresses the word "account".

Every time Gilbert replies affirmatively ("yes"), or turns the floor over to Gilbert ("all right"), he nods his head, and raises his eyebrows. George and Gilbert look at each other when Gilbert asks a question, but at the end of each question, Gilbert looks up slightly. During the brief pause at the end of affirmative statements the speaker (always George, in this fragment) blinks. To mark the end of the questions, Gilbert raises his eyebrows.

In saying the word "check", Gilbert sketches the outlines of a check in the air between him and his listener. In saying "account", Gilbert forms a kind of box in front of him with his hands: a metaphorical representation of a bank account in which one keeps money. When he says the phrase "withdraw fifty dollars," Gilbert withdraws his hand towards his chest.

1.5 Communicative Significance of the Face

Movements of the head and facial expressions can be characterized by their placement with respect to the linguistic utterance and their significance in transmitting information [61, 19, 27, 37]. The set of facial movement clusters contains: *syntactic functions* accompany the flow of speech and are synchronized at the verbal level. Facial movement can appear on accented syllable or a pause (like raising the eyebrows while saying "Do you have a blank CHECK?"); *semantic functions* can emphasize what is being said, substitute for a word or refer to an emotion (like smiling when remembering a happy event "It was such a NICE DAY"); *dialogic functions* regulate the flow of speech and depend on the relationship between two people (smooth turns[1] are often co-occurrent with mutual gaze). These three functions are modulated by various parameters: *speaker and listener characteristic functions* convey information about the speaker's social identity, emotion, attitude, age and *listener functions* correspond to the listener's reactions to the speaker's speech; they can be signals of agreement, of attention, of comprehension.

1.6 Communicative Significance of Hand Gestures

We have taken McNeil's taxonomy [45] of gesture as a working hypothesis. According to this scheme, there are four basic types of gestures during speaking [45]. *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?". The system does not make any distinction between these three categories of gesture. They are all in effect lexical. 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

[1] Meaning that the listener does not interrupt or overlap the speaker.

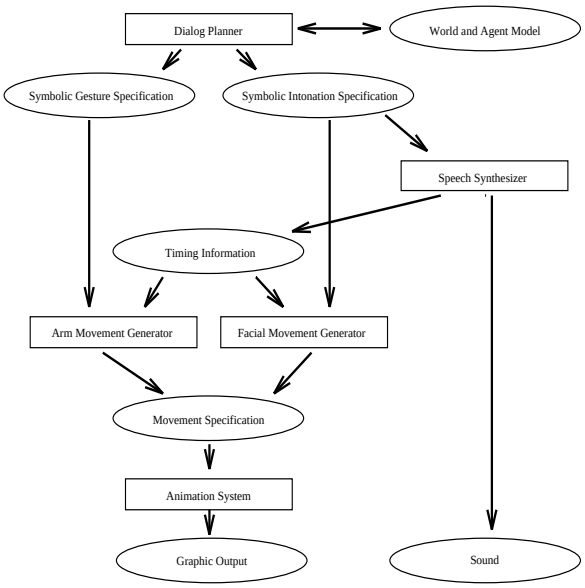

Figure 1: Architecture of each conversational agent

linguistic work. An example is waving one's left hand briefly up and down along with the stressed words in the phrase "Go AHEAD."

1.7 Synchrony of Gesture, Facial Movements, and Speech

Speech, gesture, facial expressions and gaze are intimately linked. We will assume as a working hypothesis that gestures are generated in synchrony with their semantically parallel linguistic units although in cases of hesitations, pauses or syntactically complex speech, it is the gesture which appears first [45]. The empirical basis for this assumption remains to be established (in future work, in fact, we regard our system partly as a tool for carrying out such investigations.) Hand gesture pattern and gaze pattern interfere with each other [5]. Depending on their functions, facial movements are synchronized at the phonemic segment, word or utterance levels [14, 36]. Facial expression, eye gaze and hand gestures do not do their communicative work only within single utterances, but also have inter-speaker effects. The presence or absence of confirmatory feedback by one conversational participant, via gaze or head movement, for example, affects the behavior of the other. A conversation consists of the exchange of meaningful utterances and of behaviors. One person punctuates and reinforces her speech by head nods, smiles, and hand gestures; the other person can smile back, vocalize, or shift gaze to show participation in the conversation.

2 Overview of System

In our implemented system, we have attempted to adhere as closely as possible to a model of face-to-face interaction suggested by the results of empirical research described above. In particular, each agent in conversation is implemented as an autonomous construct that maintains its own representations of the state of the world and the conversation, and whose behavior is determined by these representations. (For now, the two agents run copies of the same program, initialized with different goals and world knowledge.) The agents communicate with one another only by the symbolic messages whose content is displayed in the resulting animation. The architecture of a conversational agent is shown in Figure 1.

In this section, we provide an outline of how each agent decides what to say, determines the contribution of this content to the conversation, and uses the resulting representations to accompany speech with contextually appropriate intonation, gesture, facial expression and gaze.

2.1 Dialogue Planner

The selection of content for the dialogue by an agent is performed by two cascaded planners. The first is the domain planner, which manages the plans governing the concrete actions which an agent will execute; the second is the discourse planner, which manages the communicative actions an agent 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 an agent, and the beliefs of the agent about the world, including the beliefs of the agent about the other agent in the conversation. 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. Once decomposition resolves a plan into a sequence of actions to performed, the domain planner causes an agent to execute those actions in sequence. As these goal expansions and action executions take place, the domain planner also dictates discourse goals that an agent 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.

Following Halliday [29] and others [28, 43, 7, 63] , we use the terms *theme* and *rheme* to denote two distinct information structural attributes of an utterance.[2] The theme roughly corresponds to what the utterance is about or the question under discussion. The rheme corresponds to what the speaker has to contribute on that theme. 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. 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) [58].

Distinct intonational tunes have been shown to be associated with the thematic and rhematic parts of an utterance for certain classes of dialogue [55, 56, 57, 63]. In particular, we note that the standard rise-fall intonation generally occurs with the rhematic part of many types of utterances. Thematic elements of an utterance are often marked by a rise-fall-rise intonation.

Text is generated and pitch accents and phrasal melodies are placed on generated text as outlined in [63] and [55]. This text is converted automatically to a form suitable for input to the AT&T Bell Laboratories TTS synthesizer [41] [3]. When the dialogue is generated, the following information is saved automatically: (1) the timing of the phonemes and pauses, (2) the type and place of the accents, (3) the type and place of the gestures.

This speech and timing information is critical for synchronizing the facial and gestural animation.

2.2 Symbolic Gesture Specification

The dialogue generation program annotates utterances according to how their semantic content could relate to a spatial expression (literally, metaphorically, spatializeably, or not at all). Further, references to entities are classified according to discourse status as either new to discourse and hearer (indefinites), new to discourse but not to hearer (definites on first mention), or old (all others) [58]. According to the following rules, these annotations, together with the earlier ones, determine which concepts will have an associated gesture. Gestures that represent something (iconics and metaphorics) are generated for rhematic verbal elements (roughly, information not yet spoken about) and for hearer new references, provided that the semantic content is of an appropriate class to receive such a gesture: words with literally spatial (or concrete) content get iconics (e.g. "check" as in "do you have a blank check? or "write" as in "you can write the check", first and third frames in Figure 2); those with metaphorically spatial (or abstract) content get metaphorics (e.g. "help" as in "will you help me?", second frame in Figure 2); words with physically spatializeable content get deictics (e.g. "this bank"). All of this is done by lexical lookup. Beat gestures are on the other hand generated for such items when the semantic content cannot be represented spatially, and are also produced accompanying discourse new definite references

[2] 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*.

[3] We suppressed TTS default intonation assignment algorithm.

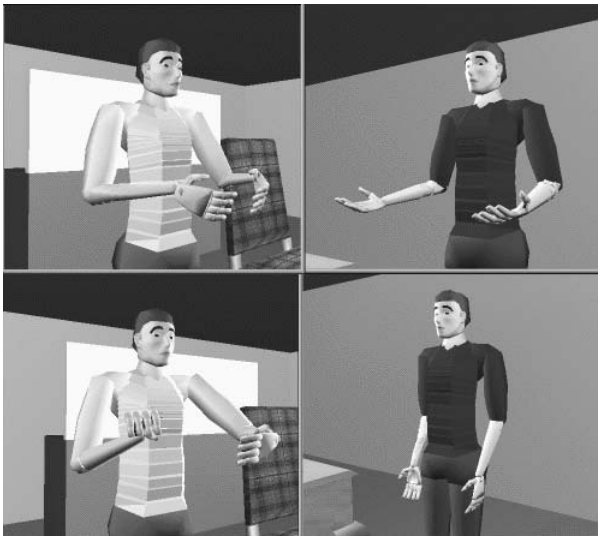

Figure 2: Examples of symbolic gesture specification

(e.g. "wait for" as in "I will wait for you to withdraw fifty dollars", fourth frame in Figure 2). If a representational gesture is called for, the system accesses a dictionary of gestures (motion prototypes) that associates semantic representations with possible gestures that might represent them[4] (for further details, see [9]).

After this gestural annotation of all gesture types, and lexicon look-up of appropriate forms for representational gestures, information about the duration of intonational phrases (acquired in speech generation) is used to time gestures. First, all the gestures in each intonational phrase are collected. Because of the relationship between accenting and gesturing, in this dialogue at most one representational gesture occurs in each intonational phrase. If there is a representational gesture, its preparation is set to begin at or before the beginning of the intonational phrase, and to finish at or before the next gesture in the intonational phrase or the nuclear stress of the phrase, whichever comes first. The stroke phase is then set to coincide with the nuclear stress of the phrase. Finally, the relaxation is set to begin no sooner than the end of the stroke or the end of the last beat in the intonational phrase, with the end of relaxation to occur around the end of the intonational phrase. Beats, in contrast, are simply timed to coincide with the stressed syllable of the word that realizes the associated concept. When these timing rules have been applied to each of the intonational phrases in the utterance, the output is a series of symbolic gesture types and the times at which they should be performed. These instructions are used to generate motion files that run the animation system [2].

2.3 Symbolic Facial Expression Specification

In the current system, facial expression (movement of the lips, eyebrows, etc.) is specified separately from movement of the head and eyes (gaze). In this section we discuss facial expression, and turn to gaze in the next section.

P. Ekman and his colleagues characterize facial expressions depending on their function [18]. Many facial functions exist (such as manipulators that correspond to biological needs of the face (wetting the lips); emblems and emotional emblems that are facial expressions replacing a word, an emotion) but only some are directly linked to the intonation of the voice. In this system, facial expressions connected to intonation are automatically generated, while other kinds of expressions (emblems, for example) are specified by hand [52].

We are using **FACS** (Facial Action Coding System [20] to define facial expressions.

[4] This solution is provisional: a richer semantics would include the features relevant for gesture generation, so that the form of the gestures could be generated algorithmically from the semantics. Note also, however, that following [37] we are led to believe that gestures may be more standardized in form than previously thought.

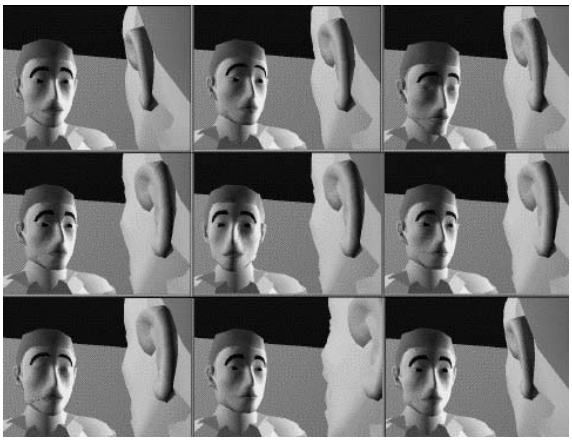

Figure 3: Facial expressions and gaze behavior corresponding to: “All right. <pause> You can write the check”.

2.4 Symbolic Gaze Specification

In the current version of the program (as in most of the relevant literature) head and eye behaviors are not differentiated. head and eyes follow the same movement pattern, and gaze is defined in terms of head motion. We identify four primary categories of gaze depending on its role in the conversation [1, 3, 13]. In the following, we give rules of action and the functions for each of these four categories (see Figure 3).

planning : corresponds to the first phase of a turn when the speaker organizes her thoughts. She has a tendency to look away (possibly in order to prevent an overload of information). On the other hand, during the execution phase, the speaker knows what she is going to say and looks more at the listener. For a short turn (duration less than 1.5 sec.), the speaker and the listener establish eye contact (mutual gaze) [1].

comment : accompanies and comments speech, by occurring in parallel with accent and emphasis [25]. Accented or emphasized items are punctuated by head nods; the speaker looks toward the listener. The speaker also gazes at the listener more when she asks a question. She looks up at the end of the question.

control : controls the communication channel and functions as a synchronization signal: responses may be demanded or suppressed by looking at the listener. When the speaker wants to give her turn of speaking to the listener, she gazes at the listener at the end of the utterance. When the listener asks for the turn, she looks up at the speaker [26].

feedback : is used to collect and seek feedback. The listener can emit different reaction signals to the speaker’s speech. Speaker looks toward the listener during grammatical pauses to obtain feedback on how utterances are being received. This is frequently followed by the listener looking at the speaker and nodding. In turn, if the speaker wants to keep her turn, she looks away from the listener. If the speaker does not emit a `within-turn` signal by gazing at the listener, the listener can still emit a `back-channel` which in turn may be followed by a `continuation signal` by the speaker. But the probability of action of the listener varies with the action of the speaker [17]; in particular, it decreases if no signal has occurred from the speaker. In this way the listener reacts to the behavior of the speaker.

3 Parallel Transition Network

Interaction between agents and synchronization of gaze and hand movements to the dialogue for each agent are accomplished using Parallel Transition Networks (PaT-Nets), which allow coordination rules to be encoded as simultaneously executing finite state automata [4]. PaT-Nets are a scheduling mechanism. They are able to take a decision, execute an action. They can call for action in the simulation and make state transitions either conditionally or probabilistically. Pat-Nets are scheduled into the simulation with an operating system that allows them to invoke or kill other PaT-Nets,

sleep until a desired time or until a desired condition is met, and synchronize with other running nets by waiting for them to finish or by waiting on a shared semaphore.

PaT-Nets are composed of nodes and transitions between these nodes. Each node corresponds to a particular state of the system. Transitions can be conditions or probabilities. When a transition is made the actions specific to the nodes are executed. In addition, the PaT-Net notation is object oriented with each net defined by a *class* with actions and transition conditions as *methods*. The running networks are instances of the PaT-Net class and can take parameters on instantiation. This notation allows PaT-Nets to be hierarchically organized and allows constructing new nets by combining existing nets or making simple modifications to existing nets.

All the behaviors specified above for gestures and gaze are implemented in PaT-Nets. First, a PaT-Net process parses the output of the speech synthesis module, one utterance at a time. Each agent has its own PaT-Net. A PaT-Net instance is created to control each agent: probabilities and other parameters appropriate for each agent given the current role as listener or speaker are set for the PaT-Net at this moment. Then as agents' PaT-Nets synchronize the agents with the dialogue and interact with the unfolding simulation they schedule activity that achieves a complex observed interaction behavior.

PaT-Nets are written in LISP. The input file is a list of lists. Each list has the form:

("George" "Gilbert" "utterance-input" 10.0)

The number 10.0 specifies the starting time of the utterance (in seconds). "George" is the speaker, "Gilbert" is the listener and "utterance-input" is the current utterance.

The PaT-Net parses each "utterance-input" and schedules actions when conditions are true. It stops when there is no more phoneme. At this point the GESTURE and GAZE PaT-Nets send information about timing and type of actions to the animation system. The animation itself is carried out by *Jack*™ [†], a program for controlling articulated objects, especially human figures.

The GAZE and GESTURE PaT-Net schedule motions as they are necessary given the current context in semi-real time. All motions do not need to be generated in advance. The animation is performed as the input utterances are scanned. It allows for interactive control and easiness of extension of the system.

4 GESTURE PaT-Net

Upon the signaling of a particular gesture, additional PaT-Nets are instantiated; 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. The separation in two PaT-Net comes from the fact that beats are superimposed over the other types of gestures; they arise from the underlying rhythmical pulse of speaking, while other gestures arise from meaning representations. Moreover beats are free movements; they are not tied to a particular gestures like the iconic or emblems movements. Beat-Net takes this freedom in consideration when computing beats gestures.

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 following example illustrates how gesture PaT-Net acts on a given input. Intonation and gesture streams are specified.

Intonation: do you have@lstar a blank@lstar check@lstarhhs
Gesture: do@st(icon-check-gesture) you have@btpr a blank@bt check@skrxed

'Check' receives a pitch accent ('lstar' stands for low pitch accent) and 'blank' a secondary stress. The beat gesture ('bt') begins its preparation phase ('btpr') on 'have'. Its stroke ('sk') falls on 'check' and the relaxation phase ('rx') starts afterwards and ends at ('ed'). Since 'blank' receives a secondary stress, the beat gestures falls there and is superimposed over the iconic gesture.

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

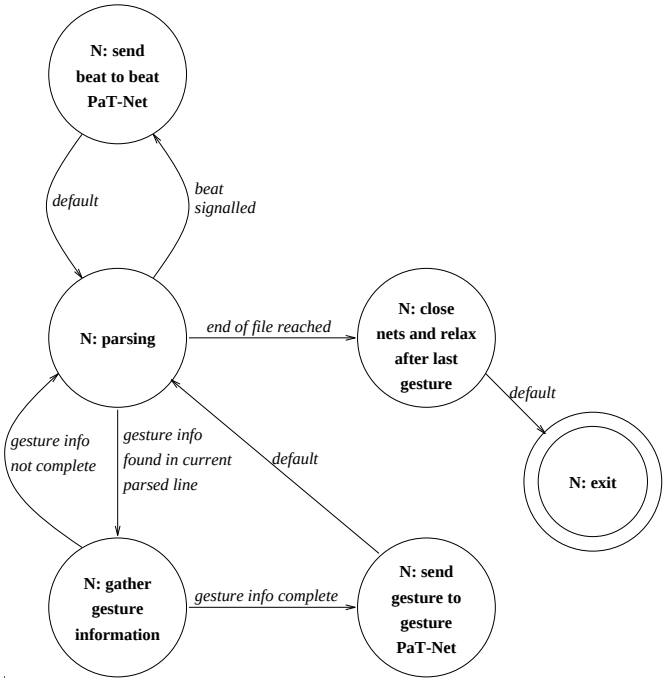

Figure 4: PaT-Net that synchronizes gestures with the dialogue at the phoneme level.

Gestures are mainly produce by speakers and not by listeners [45]. The timing of a gesture follows the change of speaking turns. The end of a gesture coincides with the end of a turn. In the case a gesture does not have time to finish, it is foreshortened. Another reason for foreshortening is anticipation of the next gesture to be produced in a discourse. In anticipatory co-articulation effects, most often the relaxation phase of the foreshortened iconic, metaphoric or deictic gesture and preparation phase of the next gesture become one.

4.1 Gesture Animation

The gesture system is divided into three parts: hand shape, wrist control, and arm positioning. Each three can be specified independently of the other. The first, hand shape, relies on an extensible library of hand shape primitives based on the American Sign Language Alphabet. It allows also relaxation position. The velocity of handshape changes is limited and allows the modeling of handshape co-articulation. Thus as the speed of the gesture increases, the gestures will 'coarticulate' in a realistic manner.

The wrist motion is specified in terms of the hand direction relative to the figure (e.g. fingers forward and palm up). Joint limits allows the wrists to move in a realistic manner. Beat gestures are a specialized form of wrist motion. The direction of the movement depends on the current wrist position.

The arm motion system accepts general specifications of spatial goals and drives the arms towards those goals within the limits imposed by the arm's range of motion. The arm may be positioned by using general directions like "chest-high, slightly forward, and to the far left".

5 GAZE PaT-Net

Each of the four dialogic functions (planning, comment, control and feedback) appears as a sub-network in the PaT-Net. Each sub-network is represented by a set of nodes, a list of conditions and their associated actions.

Moreover each node of the GAZE PaT-Net is characterized by a probability. The choice of these probabilities is based on evaluation of conversation between two persons. We analyze where and when a person is gazing, smiling and/or nodding during the interaction. Each of these signals receives binary value: 1 when it happens, 0 otherwise (gaze is equal to 1 when a person is looking at the other person, 0 when looking away). The conversation is annotated

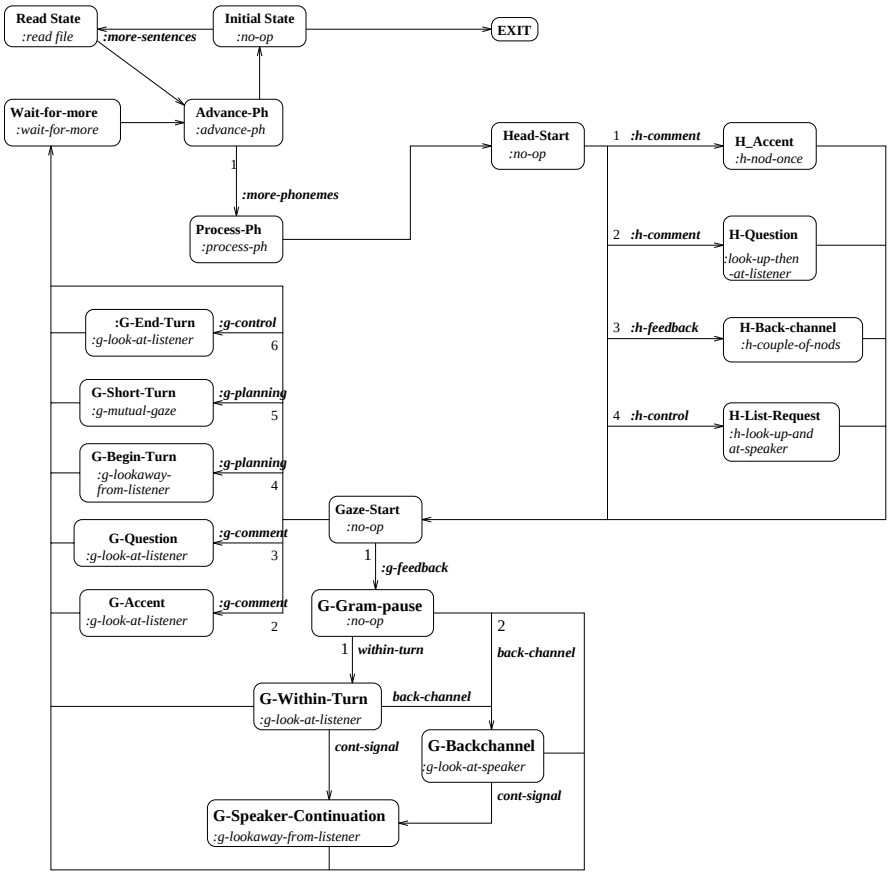

Figure 5: GAZE PaT-Net

every tenth of a second. Six turn-states are considered, three per agent. When an agent holds the floor she can be speaking while the other agent is pausing (normal turn) or speaking (overlapping talk or occurrence of backchannel), or she can be pausing as well as the listener. For each of these turn-states we compute the co-occurrence of signals: nod, smile and/or gaze. Then probability of occurrence is computed: each node of the PaT-Net corresponds to a particular configuration of a given turn-state and an occurrence of a given signal. E.g. the occurrence of a "within-turn signal" as we defined it corresponds to the action: agent1 looks at the agent2 while having the floor and pausing. Probabilities appropriate for each agent given the current role as listener or speaker are set for the PaT-Net before it executes.

Figure 5 shows the GAZE PaT-Net configuration. Node names are written in bold, actions in italic and conditions are specified on the arcs.Two sub-PaT-Nets are built: Head-Start and Gaze-Start. The differentiation refers only to the type of movements to be performed. As it can be noticed for a same condition different actions can occur. These actions are separated in two: gazing at or away from the other agent, and other head movement (nod, shake...).

Each dialogic function is associated with a set of nodes. A node is entered only if the input condition is true. The corresponding action is occurring only if its probability allows it (this check is done every time; that is when a condition is true, the action is performed only if the probability allows it. Since it is done in each case we are not repeating it). As an example the dialogic function **planning** is defined by the following nodes and conditions: if a short turn is detected then the node G-Short-Turn is entered. The associated action is: the speaker and listener look at each other. The other node corresponds to the first phase of a turn: begin-turn [5] where the speaker looks away. While the function **comment** is represented by: if an accent or a question is detected the corresponding action is executed.

[5] A beginning of a turn is defined as all the phonemes between the first one and the first accented segment.

5.1 Gaze Animation

For each phoneme, the GAZE PaT-Net is entered and decides whether or not the head should move. A transition is made on the node whose condition is true. If the probability of the node allows it, the action is performed. In such a case no further action is possible during the duration of the considered action; that is PaT-Net scheduling is disallowed and delayed until the end of this action. No co-occurrent actions are permitted. If no action is performed on a phoneme the PaT-Net waits for the next available phoneme.

Some actions performed by an agent influence the behavior of the other agent. In the case of the feedback node, different branchings are possible depending if the action "looking at the listener" is performed or not by the speaker (corresponding to a within-turn). The probability associated with a back-channel ("listener looking away from speaker") varies. It is smaller if the speaker does not look at the listener (an occurrence of a back-channel has lesser chance to happen) and greater otherwise.

Head motions are performed by tracking an invisible object. The object is placed in the current environment. The head follows the moving object. All swing, nod, turn movements are obtained by giving new coordinates to the object. A head movement is performed by giving the new position of the object, the starting time and duration of the movement. The head velocity has an easy-in/easy-out pattern, that is it accelerates at the beginning and decelerates before stopping. It allows for smoothness of movement. A nod is simulated by moving the invisible object in the vertical plane while swing and turn are executed by moving it in the horizontal plane. Each of these displacements takes as parameters the number of times to perform the move (simulation of multiple nods or swings), and distance (execution of large or small head nods). Varying these parameters allows one to use the same function for different conditions. For example, when punctuating an accent, the speaker's head nod will be of larger amplitude than the feedback head nods emitted by the listener. A gaze direction is sustained until a change is made by another action.

We illustrate a PaT-Net execution with the following example:

> `Gilbert: Get the chEck made OUt to you for fifty dollars <`*pause*`> And thEn <`*pause*`> I can withdrAw fifty dollars for you.`

planning : This utterance is not short so the node `short-turn` is not entered. But for the first few phonemes of the beginning of the example utterance (in our example "Get the ch"), the node `beginning-turn` is entered; the condition of being in a beginning of turn is true but its probability did not allow the action `speaker gazes away` to be applied. Therefore the speaker (Gilbert) keeps his current gaze direction (looking at George).

comment : In our example, on accented items ("chEck", "thEn" and "withdrAw"), the node `accent` of the function **comment** is reached. In the first two cases the probability allows the actions `speaker gazes at the listener` to be performed by Gilbert, while `nod-once` by Gilbert results on "withdraw".

control : At the end of the utterance[6] (corresponding to "fifty dollars for you" here), speaker and listener perform an action: `speaker gazes at listener` from the node `end of turn` and `listener gazes at the speaker and up` from the node `turn request`.

feedback : The two intonational phrases of our example (*get the check made out to you for fifty dollars* and *and then*) are separated by a pause; this corresponds to a within-turn situation. The node **G-feedback** is entered. If the probability allows it, the action `speaker gazes at the listener` is performed[7]. After a delay (0.2 sec., as specified by the program), the probabilities associated with the actions is checked once more. If allowed the node `back-channel` is reached, the action can happen: `listener gazes at the speaker`. In either case, the final step corresponds to the reaching of the node **speaker-continuation** after some delay. The action `speaker gazes away from the listener` is then performed.

6 Facial Expression Generation

Facial expressions are clustered into functional groups: lip shape, conversational signal, punctuator, manipulator and emblem (see Section 2.3). Each is represented by two parameters: *its time of occurrence* and *its type*. Our algorithm

[6] End of turn is defined as all the phonemes between the last accented segment and the last phonemes.

[7] In the case the action is not performed, the arc going to the node `back-channel` is immediately traversed without waiting for the next phonemic segment.

[53] embodies rules to automatically generate facial expressions, following the principle of synchrony. The program scans the input utterances and computes the different facial expressions corresponding to these functional groups.

The computation of the lip shape is done in three passes and incorporates coarticulation effects [53]. Phonemes are characterized by their degree of deformability. For each deformable segment, the program looks for the nearby segment whose associated lip shape influences it. The properties of muscle contractions are taken into account spatially (by adjusting the sequence of contracting muscles) and temporally (by noticing if a muscle has enough time to contract or relax).

A conversational signal (movements occurring on accents, like raising the eyebrow) starts and ends with the accented word; while punctuator signals (such as smiling) coincide with pauses. Blinking (occurring as punctuator or manipulator) is synchronized at the phoneme level. The extent to which the synchronic assumption holds for real human conversation is a question for future research. We hope this system and others like it will contribute to this further investigation by providing a simulation that can be controlled from the linguistic level.

7 Conclusions

Automatically generating information about intonation, facial expression, gaze, head movements and hand gestures allows an interactive dialogue animation to be created; for a non-real-time animation much guess-work in the construction of appropriate motions can be avoided. The resulting motions can be used as is, or the actions and timings can be used as a cognitively and physiologically justified guide to further refinement of the conversation and the participants' interactions by a human animator.

References

[1] M. Argyle and M. Cook. *Gaze and Mutual gaze*. Cambridge University Press, 1976.

[2] N. Badler, C. Phillips, and B. Webber. *Simulating Humans: Computer Graphics Animation and Control*. Oxford University Press, 1993.

[3] G.W. Beattie. Sequential temporal patterns of speech and gaze in dialogue. In T.A. Sebeok and J. Umiker-Sebeok, editors, *Nonverbal Communication, Interaction, and Gesture*, pages 297–320. The Hague, New-York, 1981.

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

[5] H. Bekkering, J. Pratt, and R.A. Abrams. The gap effect for eye and hand movements. *Perception and Psychophysics*, 58(4):628–635, May 1996.

[6] Alan W. Biermann, Curry I. Guinn, Richard Hipp, and Ronnie W. Smith. Efficient collaborative discourse: A theory and its implementation. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 177–181, March 1993.

[7] D. Bolinger. *Intonation and its Uses*. Stanford University Press, 1989.

[8] Tom Calvert. Composition of realistic animation sequences for multiple human figures. In Norman I. Badler, Brian A. Barsky, and David Zeltzer, editors, *Making Them Move: Mechanics, Control, and Animation of Articulated Figures*, pages 35–50. Morgan-Kaufmann, San Mateo, CA, 1991.

[9] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics Annual Conference Series*, pages 413–420, 1994.

[10] Justine Cassell, Matthew Stone, Brett Douville, Scott Prevost, Brett Achorn, Norm Badler, Mark Steedman, and Catherine Pelachaud. Modeling the interaction between speech and gesture. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, Atlanta, GA, August 1994.

[11] D. T. Chen, S. D. Pieper, S. K. Singh, J. M. Rosen, and D. Zeltzer. The virtual sailor: An implementation of interactive human body modeling. In *Proc. 1993 Virtual Reality Annual International Symposium*, Seattle, WA, September 1993. IEEE.

[12] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In M. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, Tokyo, 1993. Springer-Verlag.

[13] G. Collier. *Emotional Expression*. Lawrence Erlbaum Associates, 1985.

[14] W.S. Condon and W.D. Osgton. Speech and body motion synchrony of the speaker-hearer. In D.H. Horton and J.J. Jenkins, editors, *The Perception of Language*, pages 150–184. Academic Press, 1971.

[15] Robert Dale. *Generating Referring Expressions in a Domain of Objects and Processes*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, 1989.

[16] James Davis and Julia Hirschberg. Assigning intonational features in synthesized spoken discourse. In *ACL88*, pages 187–193, Buffalo, 1988.

[17] S. Duncan. Some signals and rules for taking speaking turns in conversations. In S. Weitz, editor, *Nonverbal Communication*. Oxford University Press, 1974.

[18] P. Ekman. Movements with precise meanings. *The Journal of Communication*, 26(3):14–26, 1976.

[19] P. Ekman. About brows: emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: claims and limits of a new disipline: contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, England; New-York, 1979.

[20] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc., 1978.

[21] I.A. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. *Proceedings of Computer Vision and Pattern Recognition (CVPR 94)*, pages 76–83, 1994.

[22] S. Feiner and K.R. McKeown. Generating coordinated multimedia explanations. In *Proceedings of the Sixth Conference on Artificial Intelligence Applications*, pages 290–296, 1990.

[23] Jean-Paul Gourret, Nadia Magnenat-Thalmann, and Daniel Thalmann. Simulation of object and human skin deformations in a grasping task. *Computer Graphics*, 23(3):21–30, 1989.

[24] Curry I. Guinn. A computational model of dialogue initiative in collaborative discourse. In *Human-Computer Collaboration: Reconciling Theory, Synthesizing Practice, Papers from the 1993 Fall Symposium Series, AAAI Technical Report FS-93-05*, 1993.

[25] U. Hadar, T.J. Steiner, E.C. Grant, and F. Clifford Rose. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129, 1983.

[26] U. Hadar, T.J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[27] U. Hadar, T.J. Steiner, and F.C. Rose. The relationship between head movements and speech dysfluencies. *Language and Speech*, 27(4):333–342, 1984.

[28] Eva Hajičová and Petr 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.

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

[30] D.R. Hill, A. Pearce, and B. Wyvill. Animating speech: An automated approach using speech synthesised by rules. *The Visual Computer*, 3:277–289, 1988.

[31] George Houghton. *The Production of Language in Dialogue: A Computational Model*. PhD thesis, University of Sussex, 1986.

[32] George Houghton and Stephen Isard. Why to speak, what to say and how to say it. In P. Morris, editor, *Modelling Cognition*. Wiley, 1987.

[33] George Houghton and M. Pearson. The production of spoken dialogue. In M. Zock and G. Sabah, editors, *Advances in Natural Language Generation: An Interdisciplinary Perspective, Vol. 1*. Pinter Publishers, London, 1988.

[34] Eduard H. Hovy. Planning coherent multisentential text. In *ACL*, pages 163–169, 1988.

[35] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. SMILE: A multilayered facial animation system. In T.L. Kunii, editor, *Modeling in Computer Graphics*. Springer-Verlag, 1991.

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

[37] Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M.R.Key, editor, *The Relation between Verbal and Nonverbal Communication*, pages 207–227. Mouton, 1980.

[38] Jintae Lee and Tosiyasu L. Kunii. Visual translation: From native language to sign language. In *Workshop on Visual Languages*, Seattle, WA, 1993. IEEE.

[39] Philip Lee, Susanna Wei, Jianmin Zhao, and Norman I. Badler. Strength guided motion. *Computer Graphics*, 24(4):253–262, 1990.

[40] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. *Computer Graphics Annual Conference Series*, pages 55–62, 1995.

[41] Mark 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.

[42] Jeffrey Loomis, Howard Poizner, Ursula Bellugi, Alynn Blakemore, and John Hollerbach. Computer graphic modeling of American Sign Language. *Computer Graphics*, 17(3):105–114, July 1983.

[43] J. Lyons. *Semantics (vol II)*. Cambridge University Press, 1977.

[44] Nadia Magnenat-Thalmann and Daniel Thalmann. Human body deformations using joint-dependent local operators and finite-element theory. In Norman I. Badler, Brian A. Barsky, and David Zeltzer, editors, *Making Them Move: Mechanics, Control, and Animation of Articulated Figures*, pages 243–262. Morgan-Kaufmann, San Mateo, CA, 1991.

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

[46] Marie W. Meteer. Bridging the generation gap between text planning and linguistic realization. *Computational Intelligence*, 7(4):296–304, 1991.

[47] Johanna D. Moore and Cécile L. Paris. Planning text for advisory dialogues. In *ACL*, pages 203–211, 1989.

[48] M. Nahas, H. Huitric, and M. Saintourens. Animation of a B-spline figure. *The Visual Computer*, 3(5):272–276, March 1988.

[49] F.I. Parke. A parameterized model for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–70, 1982.

[50] M. Patel and P.J. Willis. FACES—The facial animation, construction and editing system. In *Eurographics'91*, pages 33–45, Austria, 1991.

[51] A. Pearce, B. Wyvill, and D.R. Hill. Speech and expression: A computer solution to face animation. *Graphics and Vision Interface '86*, pages 136–140, 1986.

[52] 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.

[53] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 1996.

[54] Richard Power. The organisation of purposeful dialogues. *Linguistics*, 17(1/2):107–152, 1977.

[55] Scott Prevost and Mark Steedman. Generating contextually appropriate intonation. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, pages 332–340, Utrecht, April 1993.

[56] Scott Prevost and Mark 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, September 1993.

[57] Scott Prevost and Mark Steedman. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153, 1994.

[58] Ellen 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.

[59] Ehud Reiter. has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Seventh International Workshop on Natural Language Generation*, pages 163–170, June 1994.

[60] Hans Rijpkema and Michael Girard. Computer animation of hands and grasping. *Computer Graphics*, 25(4):339–348, July 1991.

[61] Klaus R. Scherer. The functions of nonverbal signs in conversation. In H. Giles R. St. Clair, editor, *The Social and Physiological Contexts of Language*, pages 225–243. Lawrence Erlbaum Associates, 1980.

[62] Stuart Shieber, Gertjan van Noord, Fernando Pereira, and Robert Moore. Semantic-head-driven generation. *Computational Linguistics*, 16:30–42, 1990.

[63] Mark Steedman. Structure and intonation. *Language*, 67:260–296, 1991.

[64] Akikazu Takeuchi and Katashi Nagao. Communicative facial displays as a new conversational modality. In *ACM/IFIP INTERCHI'93*, Amsterdam, 1993.

[65] Jacques Terken. The distribution of accents in instructions as a function of discourse structure. *Language and Structure*, 27:269–289, 1984.

[66] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[67] Wolfgang Wahlster, Elisabeth André, Son Bandyopadhyay, Winfried Graf, and Thomas Rist. WIP: The coordinated generation of multimodal presentations from a common representation. In Oliviero Stock, John Slack, and Andrew Ortony, editors, *Computational Theories of Communication and their Applications*. Berlin: Springer Verlag, 1991.

[68] Lyn Walker. *Informational redundancy and resource bounds in dialogue*. PhD thesis, University of Pennsylvania, 1993. Institute for Research in Cognitive Science report IRCS-93-45.

[69] R. Zacharski, A.I.C. Monaghan, D.R. Ladd, and J. Delin. BRIDGE: Basic research on intonation in dialogue generation. Technical report, HCRC: University of Edinburgh, 1993. Unpublished manuscript.

8 Research Acknowledgements

This research is partially supported by NSF Grants IRI90-18513, IRI91-17110, CISE Grant CDA88-22719, NSF graduate fellowships, NSF VPW GER-9350179; 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 Defence Science and Engineering Graduate Fellowship in Computer Science DAAL03-92-G-0342; and NSF Instrumentation and Laboratory Improvement Program Grant USE-9152503.