ANIMATED CONVERSATION:
Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents

Justine Cassell Catherine Pelachaud Norman Badler Mark Steedman
Brett Achorn Tripp Becket Brett Douville Scott Prevost Matthew Stone [1]
Department of Computer & Information Science, University of Pennsylvania

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. Coordinated arm, wrist, and hand motions are invoked to create semantically meaningful gestures. Throughout, we will use examples from an actual synthesized, fully animated conversation.

1 Introduction

When faced with the task of bringing to life a human-like character, few options are currently available. Either one can manually and laboriously manipulate the numerous degrees of freedom in a synthetic figure, one can write or acquire increasingly sophisticated motion generation software such as inverse kinematics and dynamics, or one can resort to "performance-based" motions obtained from a live actor or puppet. The emergence of low-cost, real-time motion sensing devices has led to renewed interest in active motion capture since 3D position and orientation trajectories may be acquired directly rather than from tedious image rotoscoping [34]. Both facial and gestural motions are efficiently tracked from a suitably harnessed actor. But this does not imply that the end of manual or synthesized animation is near. Instead it raises the challenge of providing a sophisticated toolkit for human character animation that does not require the presence nor skill of a live actor [2], thus freeing the skilled animator for more challenging tasks.

In this paper we present our system for *automatically animating conversations between multiple human-like agents with appropriate and synchronized speech, intonation, facial expressions, and hand gestures*. Especially noteworthy is the linkage between speech and gesture which has not been explored before in synthesizing realistic animation. In people, speech, facial expressions, and gestures are physiologically linked. While an expert animator may realize this unconsciously in the "look" of a properly animated character, a program to automatically generate motions must know the rules in advance. This paper presents a working system to realize interacting animated agents.

Conversation is an interactive dialogue between two agents. Conversation includes 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 all of these verbal and non-verbal behaviors, one cannot have realistic, or at least believable, autonomous agents. To limit the problems (such as voice and face recognition) that arise from the involvement of real human conversants, and to constrain the dialogue, we present the work in the form of a dialogue generation program in which two copies of an identical program having different knowledge of the world must cooperate to accomplish a goal. Both agents of the conversation collaborate via the dialogue to develop a simple plan of action. They interact with each other to exchange information and ask questions.

In this paper, we first present the background information necessary to establish the synchrony of speech, facial expression, and gesture. We then discuss the system architecture and its several subcomponents.

2 Background

Faces change expressions continuously, and many of these changes are synchronized to what is going on in concurrent conversation. Facial expressions are linked to the content of speech (scrunching one's nose when talking about something unpleasant), emotion (wrinkling one's eyebrows with worry), personality (frowning all the time), and other behavioral variables. Facial expressions can replace sequences of words ("she was dressed [wrinkle nose, stick out tongue]") as well as accompany them [16], and they can serve to help disambiguate what is being said when the acoustic signal is degraded. They do not occur randomly but rather are synchronized to one's own speech, or to the speech of others [13], [20].

Eye gaze is also an important feature of non-verbal communicative behaviors. Its main functions are to help regulate the flow of conversation, signal the search for feedback during an interaction (gazing at the other person to see how she follows), look for information, express emotion (looking downward in case of sadness), or influence another person's behavior (staring at a person to show power)[14].

People also produce hand gestures spontaneously while they speak, and such gestures support and expand on information con-

[1]The authors would like to thank Francisco Azuola, Chin Seah, John Granieri, Ioi Kim Lam, and Xinmin Zhao.

veyed by words. The fact that gestures occur at the same time as speech, and that they carry the same meaning as speech, suggests that the production of the two are intimately linked. In fact, not only are the meaning of words and of gestures intimately linked in a discourse, but so are their functions in accomplishing conversational work: it has been shown that certain kinds of gestures produced during conversation act to structure the contributions of the two participants (to signal when an utterance continues the same topic or strikes out in a new direction), and to signal the contribution of particular utterances to the current discourse. It is clear that, like facial expression, gesture is not a kinesic performance independent of speech, or simply a 'translation' of speech. Rather, gesture and speech are so intimately connected that one cannot say which one is dependent on the other. Both can be claimed to arise from a single internal encoding process ([8], [21], [27]).

2.1 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 intersynchronization 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 humans 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 4.1).

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.

2.2 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 [35]. The set of facial movement clusters contains:

- *syntactic functions* accompany the flow of speech and are synchronized at the verbal level. Facial movements (such as raising the eyebrows, nodding the head or blinking while saying "do you have a blank CHECK") can appear on an accented syllable or a pause.
- *semantic functions* can emphasize what is being said, substitute for a word or refer to an emotion (like wrinkling the nose while talking about something disgusting or smiling while 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; e.g at the end of "do you have a blank check?", both interactants look at each other).

These three functions are modulated by various parameters:

- *speaker and listener characteristic functions* convey information about the speaker's social identity, emotion, attitude, age (friends spend more time looking at each other while talking than a lying speaker who will avoid the other's gaze).
- *listener functions* correspond to the listener's reactions to the speaker's speech; they can be signals of agreement, of attention, of comprehension (like saying "I see", "mhmm").

2.3 Communicative Significance of Hand Gestures

Gesture too can be described in terms of its intrinsic relationship to speech. Three aspects of this relationship are described before we go on to speak about the synchronization of the two communicative channels.

First of all, four basic types of gestures occur only during speech ([27] estimates that 90% of all gestures occur when the speaker is actually uttering something).

- *Iconics* represent some feature of the accompanying speech, such as sketching a small rectangular space with one's two hands while saying "do you have a blank CHECK?"
- *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 "then I can WITHDRAW fifty dollars for you".
- *Deictics* indicate a point in space. They accompany reference to persons, places and other spatializeable discourse entities. An example might be pointing to the ground while saying "do you have an account at THIS bank?".
- *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 phrase "all right".

.

In some discourse contexts about three-quarters of all clauses are accompanied by gestures of one kind or another; of these, about 40% are iconic, 40% are beats, and the remaining 20% are divided between deictic and metaphoric gestures [27]. And surprisingly, although the proportion of different gestures may change, all of these types of gestures, and spontaneous gesturing in general, are found in discourses by speakers of most languages.

There is also a 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. For example, one speaker, describing how one deposits checks into a bank account, said "you list the checks" while she depicted with her hands that the deposit slip is to be turned over and turned vertically in order for the checks to be listed in the spaces provided on the back of the slip.

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

Finally, the importance of the interdependence of speech and gesture is shown by the fact that speakers rely on information conveyed in gesture – sometimes even to the exclusion of information conveyed by accompanying speech – as they try to comprehend a story [9].

Nevertheless, hand gestures and gaze behavior have been virtually absent from attempts to animate semi-autonomous agents in communicative contexts.

2.4 Synchrony of Gesture, Facial Movements, and Speech

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 behavior. 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.

Synchrony implies that changes occurring in speech and in body movements should appear at the same time. For example, when a word begins to be articulated, eye blinks, hand movement, head turning, and brow raising can occur and can finish at the end of the word.

Synchrony occurs at all levels of speech: the phonemic segment, word, phrase or long utterance. Different facial motions are characteristic of these different groups [13], [20]. Some of them are more adapted to the phoneme level, like an eye blink, while others act at the word level, like a frown. In the example "Do you have a blank check?", a raising eyebrow starts and ends on the accented syllables "check", while a blink starts and ends on the pause marking the end of the utterance. Facial expression of emphasis can match the emphasized segment, showing synchronization at this level (a sequence of head nods can punctuate the emphasis). Moreover, some movements reflect encoding-decoding difficulties and therefore coincide with hesitations and pauses inside clauses. Many hesitation pauses are produced at the beginning of speech and correlate with avoidance of gaze (the head of the speaker turns away from the listener) as if to help the speaker to concentrate on what she is going to say.

Gestures occur 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 ([27]). At the most local level, individual gestures and words are synchronized in time so that the 'stroke' (most energetic part of the gesture) occurs either with or just before the phonologically most prominent syllable of the accompanying speech segment ([21], [27]). At the most global level, we find that the hands of the speaker come to rest at the end of a speaking turn, before the next speaker begins her turn. At the intermediate level, the phenomenon of co-articulation of gestural units is found, whereby gestures are performed rapidly, or their production is stretched out over time, so as to synchronize with preceding and following gestures, and the speech these gestures accompany. An example of gestural co-articulation is the relationship between the two gestures in the phrase "get the check MADE OUT TO YOU for fifty dollars and then I can WITHDRAW fifty dollars for you". During the phrase 'made out to you', the right hand sketches a writing gesture in front of the speaker. However, rather than carrying this gesture all the way to completion (either both hands coming to rest at the end of this gesture, or maintaining the location of the hands in space), the hand drops slightly and then pulls back towards the speaker to perform the 'withdraw' gesture. Thus, the occurrence of the phrase 'made out to you', with its accompanying gesture, affected the occurrence of the gesture that accompanied "withdraw".

3 Computer Animation of Conversation

3.1 Literature on Facial Control Systems

Various systems have been proposed to integrate the different facial expression functions. Most of the systems use **FACS** (Facial Action Coding System) as a notational system [17]. This system is based on anatomical studies, and describes any visible facial movements. An action unit **AU**, the basic element of this system, describes the action produced by one or a group of related muscles.

The multi-layer approach [19] allows independent control at each level of the system. At the lowest level (geometric level), geometry of the face can be modified using free form deformation techniques. At the highest level, facial animation can be computed from an input utterance.

In M. Patel's model [28] facial animation can also be done at different levels of representation. It can be done either at the muscle level, the **AU** level or the script level. For each **AU** the user can select starting and ending points of action, the intensity of action, the start and end tensions and the interpolation method to compute the in-between frames. An alternative approach is proposed by [11] with good results.

Building a user-interface, [37] propose a categorization of facial expressions depending on their communicative meaning. For each of the facial functions a list of facial displays is performed (for example, remembering corresponds to eyebrow action, eye closure and one side of mouth pull back). A user talks to the 3D synthetic actor. A speech system recognizes the words and generates an answer with the appropriate facial displays. Grammar rules, a small vocabulary set and a specific knowledge domain are part of the speech analysis system. The responses by the 3D actor are selected from a pre-established set of utterances. The appropriate facial displays accompanying the answer follow the analysis of the conventional situation (e.g. if the user's speech is not recognized the 3D actor will answer with a "not-confident" facial display).

3.2 Literature on Gesture Animation

The computer graphics literature is rather sparse on the topic of gesture animation. Animators frequently use key parameter techniques to create arm and hand motions. Rijpkema and Girard [33] created handshapes automatically based on the object being gripped. The Thalmanns [18, 26] improved on the hand model to include much better skin models and deformations of the finger tips and the gripped object. Lee and Kunii [22] 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 [25]. Chen et al [10] constructed a virtual human that can shake hands with an interactive participant. Lee et al [23] automatically generate lifting gestures by considering strength and comfort measures. Moravec and Calvert [5] 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.

4 Overview of System

In the current system, a model of face-to-face interaction is used to generate all of the behaviors implemented, from the informational status of intonation to the communicative function of head nods, gaze, and hand gestures. Additionally, however, this system implements two agents whose verbal and nonverbal behaviors are integrated not only within turns, but across speakers.

In the remaining parts of the paper we explain the different elements of Figure 1. We start from the top of the figure and work towards its bottom. Currently, gesture is generated by the dialogue

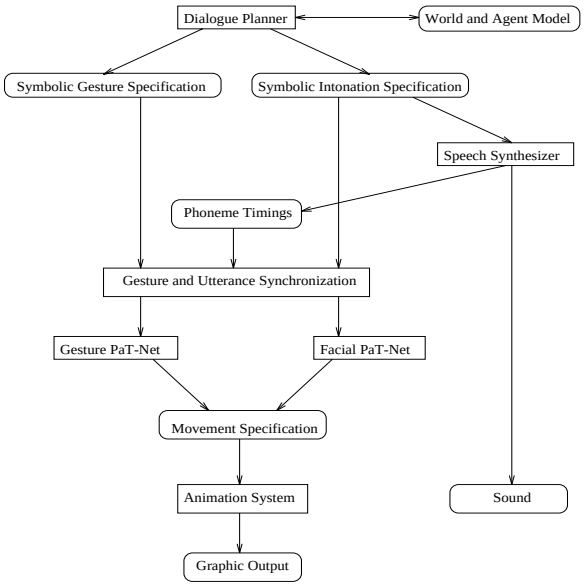

Figure 1: Interaction of components

planner, while facial expression and gaze are generated by the facial PaT-Net.

4.1 Dialogue Planner

The text of this dialogue is automatically generated on the basis of a database of facts describing the way the world works, a list of the goals of the two agents, and the set of beliefs of those two agents about the world, including the beliefs of the agents about one another [30], [7]. In this instance the two agents have goals that change over the course of the dialogue (Gilbert comes to have the goal of helping George get $50; George comes to have the goal of writing a check).

Text is generated and pitch accents and phrasal melodies are placed on generated text as outlined in [36] and [31]. This text is converted automatically to a form suitable for input to the AT&T Bell Laboratories TTS synthesizer ([24]). 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 will be critical for synchronizing the facial and gestural animation.

4.2 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) [32]. 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"); those with metaphorically spatial (or abstract) content get metaphorics (e.g. "account"); words with physically spatializeable content get deictics (e.g. "this bank"). Meanwhile, beat gestures are generated for such items when the semantic content cannot be represented spatially, and are also produced accompanying discourse new definite references (e.g. "fifty dollars"). 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[2] (for further details, see [7]).

Figure 2: Examples of symbolic gesture specification

In Figure 2, we see examples of how symbolic gestures are generated from discourse content.

1. "Do you have a BLANK CHECK?"
   - In the first frame, an iconic gesture (representing a rectangular check) is generated from the first mention (new to hearer) of the entity 'blank check'.
2. "Will you HELP me get fifty dollars?"
   - In the second frame, a metaphoric gesture (the common *propose* gesture, representing the request for help as a proposal that can be offered to the listener) is generated because of the first mention (new to hearer) of the request for help.
3. "You can WRITE the check."
   - In the third 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'.
4. "I will WAIT for you to withdraw fifty dollars for me."
   - In the fourth frame, a beat gesture (a movement of the hand up and down) is generated from the first mention of the notion 'wait for', which cannot be represented spatially.

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

[2] 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 [21] we are led to believe that gestures may be more standardized in form than previously thought.

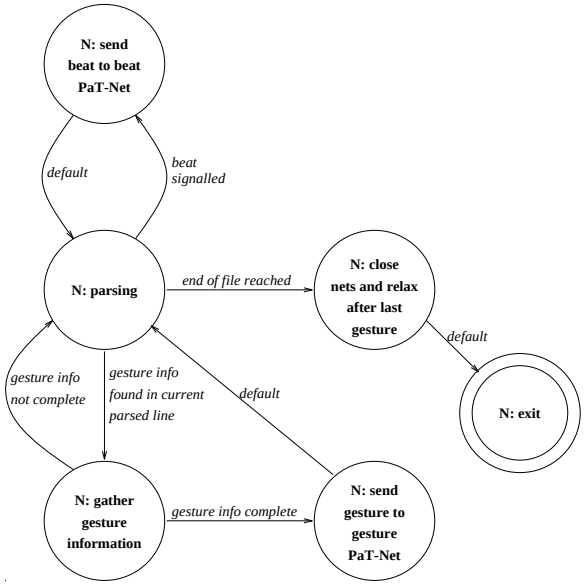

Figure 3: Pat-Net that synchronizes gestures with the dialogue at the phoneme level.

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 ([3]).

4.3 The Underlying Coordination Model

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 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.

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.

Behaviors are implemented as specified in the following sections, with all head, eye and hand movement behavior for an individual encoded in PaT-Nets. A PaT-Net instance is created to control each agent with appropriate parameters. 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.

4.4 Gesture Generator

The gesture PaT-Net sends information about the timing, shape, and position of the hands and arms to the animation system. The animation process produces a file of motions to be carried out by the two figures. Starting with a given gesture and its timing, speech rate and surrounding gestures constrain the motion sequence for a proper co-articulation effect. As depicted in Figure 3, 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" ([38]) 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.

Consider the following data from the intonation and gesture streams. Let us examine a gesture PaT-Net that acts on this input.
Intonation: Do you have a blank CHECK
Gesture: pr beat sk rx
In this example, the primary intonational stress of the phrase falls on 'check', but there is a secondary stress on 'blank'. The gesture line of the example shows that the preparation ('pr') of the gesture begins on 'have', that the stroke of the gesture ('st') falls on check, and that the gesturing relaxes ('rx') after 'check'. Because of the secondary stress on the new informational item 'blank', a beat gesture falls there, and it is found superimposed over the production of the iconic gesture.

Due to the structure of the conversation, where the speakers alternate turns, we assume similar alternation in gesturing. (Gesturing by listeners is almost non-existent [27].) For the purposes of gesture generation, phoneme information is ignored; however, utterance barriers must be interpreted both to provide an envelope for the timing of a particular gesture or sequence of gestures and to determine which speaker is gesturing. Timing information, given in the speech file, also allows the PaT-Net to determine whether there is enough time for a complete gesture to be produced. For example, the iconic gesture which accompanies the utterance "*Do you have a blank [check]?*" has sufficient time to execute: it is the only (non-beat) gesture occurring in the phrase, as shown above. However, if this timing is insufficient to allow for full gesture production, then the gesture must be foreshortened to allow for the reduced available timing (because beat gestures are produced by a separate PaT-Net system, they do not enter into questions of co-articulation).

The most common 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. This process can be seen in the gestures accompanying the phrase "*Get the check [made out to you] for fifty dollars and then I can [withdraw] fifty dollars for you*". "[Made out to you]" is produced .90 seconds into the phrase, and

"[withdraw]" is generated at 1.9 seconds. This causes some foreshortening in the relaxation process during the first gesture, from which the second gesture is then produced.

Co-articulation constraints – synchronizing the gestures with intonational phrases and surrounding gestures – may actually cause the given gestures to be aborted if too little time is available for production given the physical constraints of the human model.

4.5 Gesture Motion Specification

The graphics-level gesture animation system accepts gesture instructions containing information about the location, type, timing, and handshape of individual gestures. Based on the current location of the hands and arms in space, the system will attempt to get as close as possible to the gesture goals in the time allowed, but may mute motions or positionings because it cannot achieve them in time (co-articulation effects). This animation system calls upon a library of predefined handshapes which form the primitives of hand gesture. These handshapes were chosen to reflect the shapes most often found in gesture during conversational interaction ([21]). The animation system also calls upon separate hand, arm and wrist control mechanisms.

The gesture system is divided into three parts: hand shape, wrist control, and arm positioning. The first, hand shape, relies on an extensible library of hand shape primitives for the basic joint positions, but allows varying degrees of relaxation towards a neutral hand position. The speed at which the hand may change shape is also limited to allow the modelling of hand shape co-articulation. Large changes in hand position are restricted as less time is allotted for the hand movement, forcing faster hand gestures to smooth together.

The wrist control system allows the wrist to maintain and change its position independently of what complex arm motions may be occurring. The wrist is limited within the model to a physically realistic range of motion. Wrist direction is specified in terms of simple directions relative to the gesturer, such as "point the fingers of the left hand forward and up, and the palm right".

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".

The expressiveness of an individual's gesturing can be represented by adjusting the size of the gesture space of the graphical figure. In this way, parameters such as age (children's gestures are larger than adults') and culture (in some cultures gestures tend to be larger) can be implemented in the gesture animation.

4.6 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 the set of semantic and syntactic facial expressions depending on their meaning [15]. 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 [29].

4.7 Symbolic Gaze Specification

Gaze can be classified into four primary categories depending on its role in the conversation [1], [12]. In the following, we give rules of action and the functions for each of these four categories (see Figure 4). The nodes of the Pat-Net they refer to is also indicated.

Figure 4: Facial expressions and gaze behavior corresponding to: "All right. <pause> You can write the check".

planning: corresponds to the first phase of a turn when the speaker organizes her thoughts. She has a tendency to look away in order to prevent an overload of information (`beginning of turn`). 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] (`short-turn`).

comment: accompanies and comments speech, by occurring in parallel with accent and emphasis. Accented or emphasized items are punctuated by head nods; the speaker looks toward the listener (`accent`). The speaker also gazes at the listener more when she asks a question. She looks up at the end of the question (`utterance: question`). When answering, the speaker looks away (`utterance: answer`).

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 (`end of turn`). When the listener asks for the turn, she looks up at the speaker (`turn request`).

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 (`within-turn`). This is frequently followed by the listener looking at the speaker and nodding (`back-channel`). In turn, if the speaker wants to keep her turn, she looks away from the listener (`continuation signal`). If the speaker doesn't 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 [14]; in particular, it decreases if no signal has occurred from the speaker. In this way the listener reacts to the behavior of the speaker.

4.8 Gaze Generator

Each of the dialogic functions appears as a sub-network in the PaT-Net. Figure 5 outlines the high-level PaT-Net for gaze control for a single agent. It contains the four dialogic functions, their nodes that define each function, and their associated actions. From the definitions given above, we extract the conditions and the actions characterizing the dialogic functions. For this current version of the program we do not differentiate head movement and eye movement.

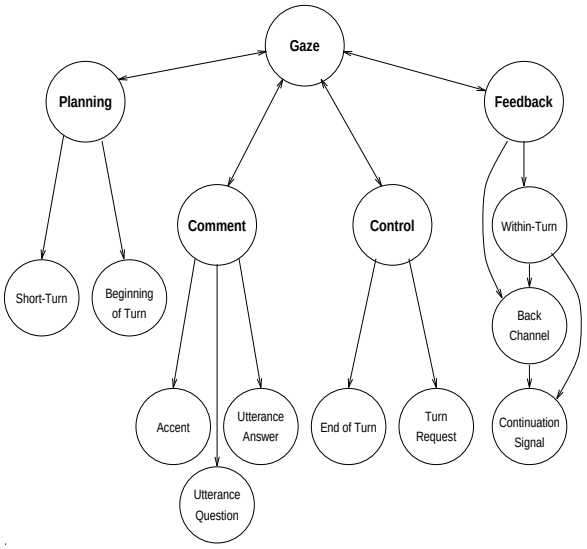

Figure 5: The gaze movement PaT-Net: actions are defined in the nodes; conditional and probabilistic transitions occur on arcs. All leaf nodes also branch back to the root node unconditionally.

That is the eyes follow the head. Moreover, in the literature this difference is rarely made. In what follows, we use "gaze" to refer to head and eye movement.

Each node is characterized by a probability. A person can have the floor talking or pausing, but loses it as soon as the other person starts talking. There are 3 possible states per person while having the floor. If Speaker has the floor: Speaker talks and Listener pauses, both of them are talking, or both of them are pausing. For each of these states, Speaker and Listener can gaze at each other or not. This gives us 12 possibilities, or 24 per dyad. We can then compute the probability of being in each of these states [6]. Most of the nodes of the Pat-Net can be characterized by a certain set of states. For example the occurrence of a "within-turn signal" as we defined it corresponds to the action: person1 looks at the person2 while having the floor and pausing. These state sets correspond to a sub-matrix. We compute the probability of each sub-matrix in relation to the particular state (having the floor and pausing) to arrive at a probability of occurrence. We do such a computation for all the other nodes of the Pat-Net. Probabilities appropriate for each agent given the current role as listener or speaker are set for the PaT-Net before it executes. At each turn change, the probabilities change values accordingly. This information is used to determine the rules and transitional probabilities for actions in Pat-Nets.

For each phoneme, the GAZE Pat-Net is entered. A transition is made on the node whose condition is true. If the probability of the nodes allows it, the action is performed. The action of the different nodes of the Pat-Net is illustrated in the following with the example:

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

planning: For the first few phonemes of the beginning of the example utterance [3](in our example it corresponds to "Get the ch"), the sub-network **planning** is applied. This utterance is not short so the node `short-turn` is not entered.

But 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).

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

comment: In our example, on accented items ("chEck", "thEn" and "withdrAw"), the node `accent` of the sub-network **comment** is reached; the actions `speaker gazes at the listener` and `head nod` are performed by Gilbert. As before, the instantiation of an action depends on its probability. The system easily represents the parallel agent actions.

control: In our example at the end of the utterance[4] (corresponding to "fifty dollars for you" here) the sub-network **control** is entered. Two actions are considered. The node `end of turn` corresponds to action performed by the speaker: `speaker gazes at listener`. The other node `turn request` affects the listener; the action `listener gazes at the speaker and up` is performed.

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 sub-network **feedback** is entered. If the probability allows it, the action `speaker gazes at the listener` is performed[5]. After a delay (0.2 sec., as specified by the program), the node `back-channel` is reached. Once more the program checks the probabilities associated with the actions. Two actions can happen: `listener gazes at the speaker` and/or `the listener nods`. In either case, the final step within the **feedback** sub-network is reached after some delay. The action `speaker gazes away from the listener` is then performed.

4.9 Facial Expression Generator

Facial expressions belonging to the set of semantic and syntactic functions (see section 4.6) are clustered into functional groups: lip shape, conversational signal, punctuator, manipulator and emblem. We use **FACS** to denote facial expressions. Each is represented by two parameters: *its time of occurrence* and *its type*. Our algorithm [29] embodies rules as described in Section 4.6 to automatically generate facial expressions, following the principle of synchrony.

The program scans the input utterance and computes the different facial expressions corresponding to these functional groups. The computation of the lip shape is made in three passes and incorporates coarticulation effects. Phonemes are associated to some characteristic shapes with different degree of deformability. For deformable elements, temporal and spatial constraints modify these shapes to consider their surrounding context. A conversational signal (movements occurring on accents, like the raising of an eyebrow) starts and ends with the accented word; while punctuator signal (movement occurring on pause, like frowning) happens on the pause. When a blink is one of these signals it is synchronized at the phoneme level. Other signals such as emblems and emotional emblems are performed consciously and must be specified by the user.

By varying the two parameters defining a facial expression, different speaker personalities can be obtained. For example a persuasive person can punctuate each accented word with raising eyebrows, while another person might not.

4.10 Gaze and Facial Motion Specification

The gaze directions generated in a previous stage can now be instantiated. As discussed earlier, the `GAZE` PaT-Net in Figure 5 is run for each agent at the beginning of every phoneme. Depending on the course taken through the `GAZE` network due to probabilistic branching and environmental state, the net may commit its agent to a variety of actions such as a head nod or a change in the gaze point. A change in the gaze is accomplished by supplying the human model with a 3D coordinate at which to look and a time in

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

[5]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.

which to move – the scheduled motion then begins at the current point in the simulation and has the specified duration. A head nod is accomplished by scheduling a sequence of joint motions for the neck, supplying both the angle and the angular velocity for each nod cycle. Note that the gaze controller schedules motions as they are necessary by reacting to the unfolding simulation (in fact, it does this in semi-real time) and does not have to generate all motions in advance. This makes the gaze controller easy to extend and easy to integrate with the rest of the system.

Different functions may be served by the same action, which differ only in their timing and amplitude. 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. Different head nod functions may also be characterized by varying numbers of up/down cycles. The gaze direction is sustained by calling for the agent to look at a pre-defined point in the environment until a change is made by another action.

For facial expressions, the program outputs the list of **AUs** that characterize each phonemic element and pause [29].

After scanning all the input utterances, all the actions to be performed are specified. Animation files are output. The final animation is done by combining the different output files for the gesture, face and gaze in *Jack*.

5 Conclusions

Automatically generating information about intonation, facial expression, 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 – as demonstrated in the video – 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. I. Badler, B. A. Barsky, and D. Zeltzer, editors. *Making Them Move: Mechanics, Control, and Animation of Articulated Figures*. Morgan-Kaufmann, San Mateo, CA, 1991.

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

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

[5] 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.

[6] J. Cappella. personal communication, 1993.

[7] Justine Cassell, Mark Steedman, Norm Badler, Catherine Pelachaud, Matthew Stone, Brett Douville, Scott Prevost and Brett Achorn. *Modeling the interaction between speech and gesture*. *Proceedings of the Cognitive Science Society Annual Conference*, 1994.

[8] Justine Cassell and David McNeill. Gesture and the poetics of prose. *Poetics Today*, 12:375–404, 1992.

[9] Justine Cassell, David McNeill, and Karl-Erik McCullough. Kids, don't try this at home: Experimental mismatches of speech and gesture. presented at the International Communication Association annual meeting, 1993.

[10] 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.

[11] M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In N.M. Thalmann and D.Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139-156. Springer-Verlag, 1993.

[12] G. Collier. *Emotional expression*. Lawrence Erlbaum Associates, 1985.

[13] 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.

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

[15] P. Ekman. Movements with precise meanings. *The Journal of Communication*, 26, 1976.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

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

[28] M. Patel. *Making FACES*. PhD thesis, School of Mathematical Sciences, University of Bath, Bath, AVON, UK, 1991.

[29] 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.

[30] Richard Power. The organisation of purposeful dialogues. *Linguistics*, 1977.

[31] Scott Prevost and Mark 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.

[32] 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.

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

[34] Barbara Robertson. Easy motion. *Computer Graphics World*, 16(12):33–38, December 1993.

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

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

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

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

6 Research Acknowledgments

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 Defense Science and Engineering Graduate Fellowship in Computer Science DAAL03-92-G-0342; and NSF Instrumentation and Laboratory Improvement Program Grant USE-9152503.