CONVERSATIONALLY-SMART(ER) ANIMATED AGENTS

Justine Cassell and Group, MIT Media Lab

Goal: Embodied Conversational Characters

- Autonomous self-animating characters for use in production animation, interfaces and computer games.
- Autonomy comes from underlying models of behavior and intelligence.
- Intelligence here means "social smarts"
- Social smarts is being able to engage a human in an interesting, relevant conversation with appropriate speech and body behaviors.

Motivation

- Embodied conversational characters may leverage users' natural tendencies to attribute humanness to the interface.
- Push the "face-to-face conversation" metaphor of interface to the max
- Allow communication through multiple, natural, modalities.
- Exploit graphical bodies for the kinds of intelligence they do best

A Short History of Intelligence for Animation

- Physical models
- Behavioral models
- Cognitive models
- Social & Conversational models

AI → believability → smarter

- Classical AI:
  - Knows the domain
  - Reasons about problems
- Believability:
  - Looks human
  - Engages in human-like behaviors
- Smarter:
  - Acts human (and *reacts* to humans)
  - Knows about the *function* of human-like behaviors
  - Incorporates several kinds of intelligence

Why smart is better

- We interpret all behaviors (despite ourselves).
- We attribute reactivity to all animated creatures.
- So, if behaviors are wrong, mismatched to one-another or badly timed, the agent looks stupid.
- Graphics is the only way to convey certain kinds of intelligence.
- With more intelligence in graphics, we are moving towards a face-to-face Turing test.

Some aspects of Conversational Smarts

- The same channels carry propositional information (about the content of what is being said) and interactional information (about the process of conversation).
- Propositional and interactional information are carried by verbal (speech, intonation) and visual (facial expression, gesture, posture) means.

That is . .

- Propositional Layer
  - Verbal and visual behaviors that contribute to the intended meaning.
  - Verbal: content of speech & intonation
  - Visual / Non-verbal: deictic, iconic & metaphoric gestures
- Interactional Layer
  - Verbal and visual behaviors that regulate, coordinate and manage information flow.
  - Verbal: back-channels, "uh-huh"
  - Non-verbal / visual: gaze, nods, facial expressions, etc.

An Example

- A conversation becomes increasingly synchronized (entrainment)
- Time scales vary widely -- 400 ms to 1.4 sec (multi-threadedness)
- Multi-modality follows function(gesture is there because we need it).

Some conversational behaviors

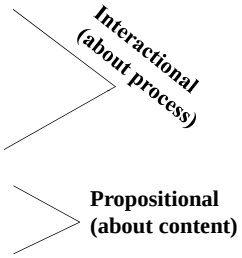
- Speech
- Intonation
- Filled pauses ("umm" & other noises)

- Eye gaze towards & away from interlocutor
- Raising eyebrows
- Nods & head shakes
- Hand gestures
- Body posture

Some functions filled by conversational behaviors

- Conversation initiation
- Giving & taking turns
- Giving and requesting feedback
- Breaking away

- Conveying information



Propositional
(about content)

Eye & Head Movement

- Eye gaze & Head turns mark
  - Status of turn-taking
  - Attention to task
  - Cognitive activity

Hand Gestures

- Mark
  - information as new and otherwise important
- Add
  - manner to description of motion
  - spatialization of people and events
  - speaker's beliefs about discourse

Human Conversation

- OK, so the name of it is Canary Row and it's got this Sylvester and Tweety guy and um what happens is it starts out and it's + um it's like a road that's separating these two buildings. One of 'em I think is a hotel - the tape kinda shoreted out for a second so you couldn't read what it was - and the other one where Sylvester is is um this thing called Bird Watcher Society, so it's kind of a joke on that, and it starts out with you seeing Sylvester. Sylvester looks, pulls out the binoculars, looks across the street at the hotel.

Iconic

- So the name of it is Canary Row
- and it's got this Sylvester and Tweety guy

Deictic

- One of ‘em I think is a hotel
- the tape kinda shorted out for a second
- so you couldn’t read what it was

Metaphoric

- Okay, so the name of it is Canary Row

Beat

- And um what happens is it starts out
- and it’s

Why do we need this stuff?

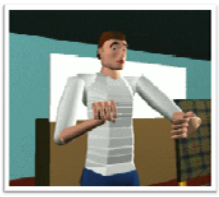
To reiterate:

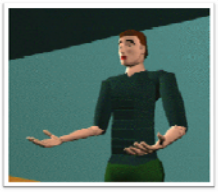
- if behaviors are wrong, or mismatched to one another, or mis-timed, the human will still assume that some meaning is intended . . . But will assume the wrong meaning.
- If behaviors are right, human will assume intelligence, credibility, interactional space.
- So we need to work from an accurate model of social-conversational skills.

How to integrate conversational smarts into Animated Agents

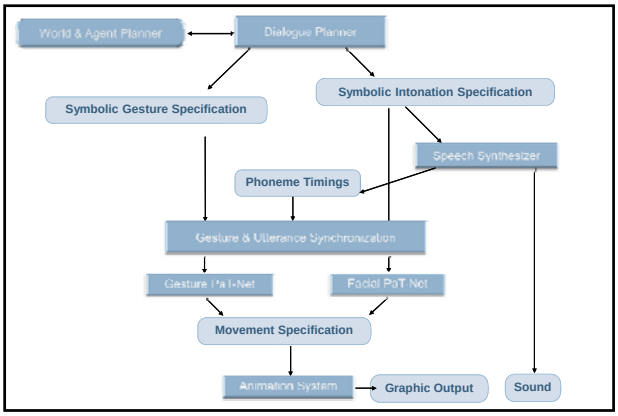
- Distinction between surface behaviors and function.
  - One behavior (a gesture) can mean several things
  - One meaning ("it's my turn") can be indicated by any one of several behaviors.
- Synchronize different channels (speech, gesture, facial expression) from early in the generation process.
- Timing, timing, timing (action scheduler)

1st Embodied Conversational Agent: Animated Conversation





See Cassell, Pelachaud, Prevost, Stone, Badler, Steedman, Deauville, 1994

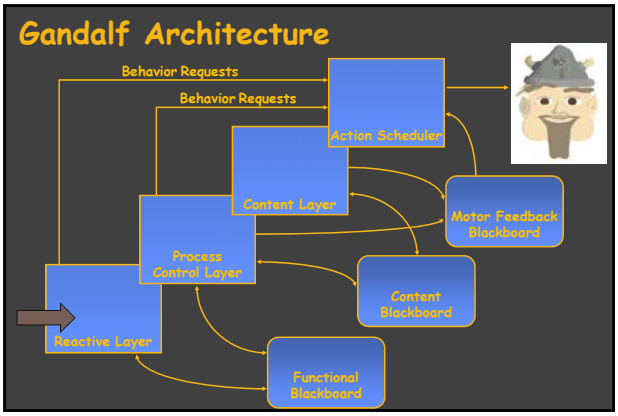


Note that

- They look like they're engaging in "foreigner talk"
- We can't help but attribute function to the number of head nods, the repetition of speech, their jerkiness.

2nd Embodied Conversational Agent: Gandalf

See Cassell & Thorisson, 1999, Thorisson, 1996



Excerpt from Gandalf

Note that

- Gandalf only knows a finite number of responses.
- Only one gesture.
- But, he interacts well enough for us to ask:
  - Do people think he's conversationally smart?

Does conversational smarts matter? Evaluation of Gandalf

- When Gandalf exhibited conversational smarts (and did not exhibit emotions), he was judged to be
  - more credible
  - more helpful
  - more collaborative

2 user studies: communicative task & collaborative task

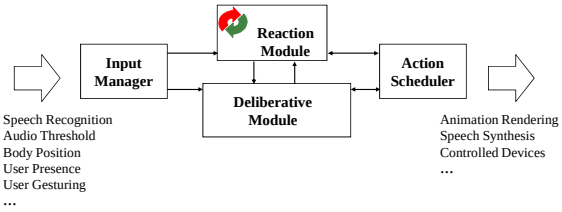
REA Embodied Conversational Agent: Case Study, in detail

- Support Multi-Modal Input and Graphical Output
- Operate in Real-Time
- Process Propositional and Interactional Information
- Use Conversational Functions (over modalities)
- Be Modular and Extensible
- Generate verbal and non-verbal output

Domain: REA, Experiment in Virtual Realty

- Shows clients through houses
- Engages in small talk
- Answers questions about particular houses
- Obeys requests to show houses/rooms
- Asks questions about client's housing needs

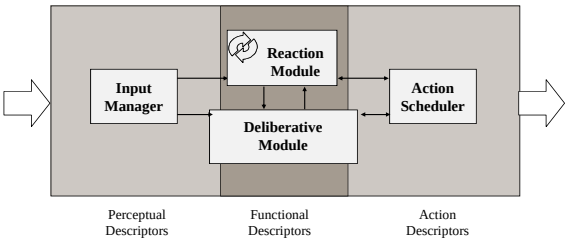
REA Architecture



REA I/O Components

- Inputs:
  - Stereo Vision: Stive
    - User present/absent, location, gesturing
    - Azarbayejani, A., Wren, C. and Pentland A. Real-time 3-D tracking of the human body. In Proceedings of IMAGE'COM 96, (Bordeaux, France, May 1996).
  - Audio threshold (speaking/paused/idle)
  - ASR: IBM ViaVoice (moving to Zue *et al.*)
- Outputs:
  - Animation: SGI OpenGL
  - TTS: Microsoft Whisper (moving to Festival)

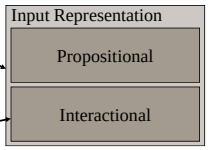
Conversational Functions



Interactional and Propositional Information

Speech
Iconic & Pointing Gesture

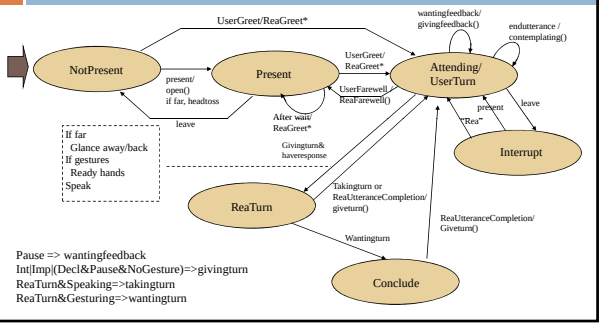
Gaze
Beat Gesture
Head Nod
Intonation
User Presence
IsGesturing
IsSpeaking



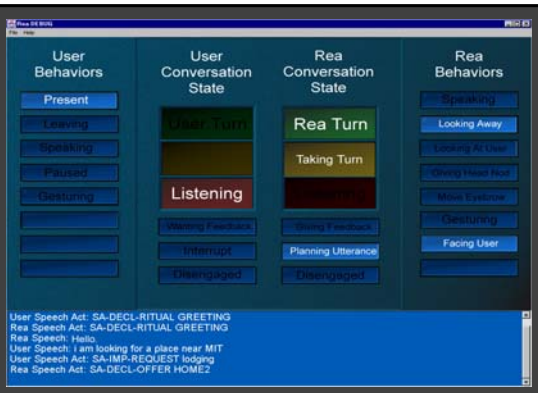
Semantic Representation

Conversational Functions
- GiveFeedback
- TakeTurn
- WantTurn
- etc.

REA Reaction Module Conversational State Diagram



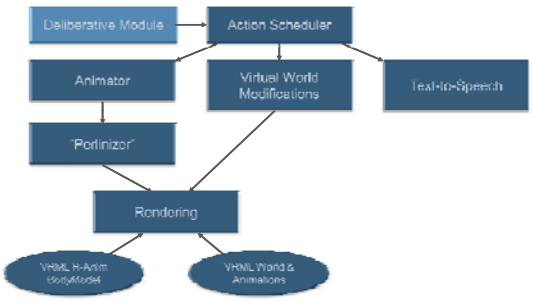
Pause => wantingfeedback
Int|Imp|(Decl&Pause&NoGesture)=>givingturn
ReaTurn&Speaking=>takingturn
ReaTurn&Gesturing=>wantingturn



REA

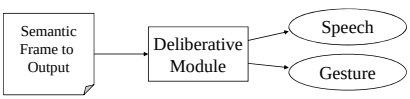
See Cassell, et al., 1999 (Proceedings of CHI)

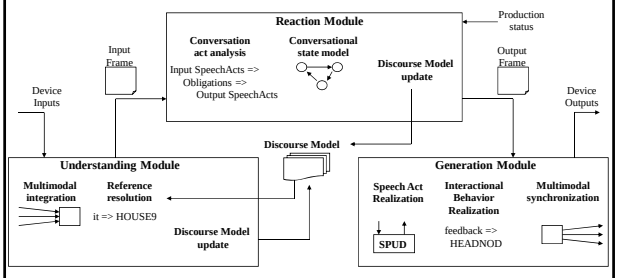
REA Output Synthesis & Coordination



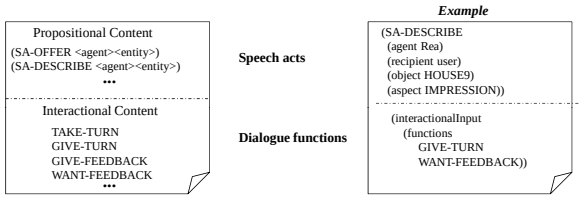
REA Deliberative Module

- Synthesizes propositional and interactional output
  - Speech and accompanying gesture
  - Decides which output modalities are best suited
  - Gestures may be complementary or redundant
  - Based on SPUD text generator (Matthew Stone)





A parenthetic note about generation of speech and gestures for interactional and propositional meanings

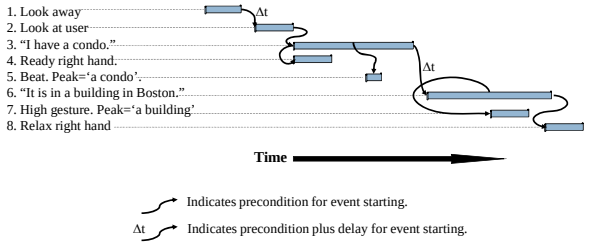


Input & Output Frame Structure

REA Action Scheduler

- Coordinates execution of multiple output channels:
  - Within a device (e.g. animated gesture & gaze)
  - Across devices (e.g., TTS and animation)
- Execution is event-driven because:
  - Very difficult to predict execution timings and start times
  - Each modality can produce events while executing an action
  - Events trigger the start or end of other actions.

Example: “I have a condo.”



Degrees of Freedom

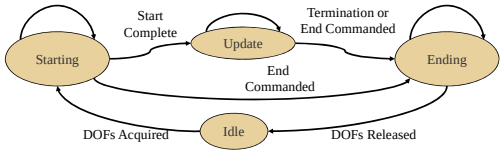
- Output resources which can be subject to contention among competing behaviors.
  - e.g., TTS, right-arm, eyes, head
- Action Scheduler handles arbitration among DOFs to ensure that only one behavior at a time can control a DOF.

Behaviors

- Represent atomic actions
  - e.g., right-arm-gesture, head-nod, speech
- Specify a set of DOFs required for execution.
- Commands:
  - start-once (execute “once” and stop)
  - start-continuous
  - stop
  - Commands issued with specified priority.

Behaviors, cont'd

- Behaviors go through three phases each time they are executed:
  - Starting and Ending *always* executed
    - e.g., Ending can return DOFs to canonical positions.



Event Rules

- Actions are specified in condition-action rules:
  - Events can be
    - immediate
    - after a specified event (or Δt after an event)
    - before a specified event
    - during a specified event
    - and/or/not combinations of the above

Action Scheduler

- Functions as
  - Rule-interpreter:
    - Determines when a rule's precondition is satisfied (or can never be satisfied)
  - Preemptive multi-tasking operating system
    - Determines if a commanded behavior can run
    - Preempts running behavior if there is a DOF conflict and running behavior has lower priority
      - The running behavior's Ending routine is always run first.

Example

(action :id **H_AWAY**	:when immediate :content (headlook :cmd away :object user))
(action :id **H_AT**	:when (offset_after :event H_AWAY.END :time 00:01.50) :content (headlook :cmd towards :object user))
(action :id **S_CONDO**	:when (after :event H_AT.END) :content (speak :content "I have a condo."))
(action	:when (after :event S_CONDO.START) :content (rightgesture :cmd ready))
(action	:when (after :event S_CONDO.WORD3) :content (rightgesture :cmd beat))
(action :id **S_BLDG**	:when (offset_after :event S_COND.END :time 00:01.00) :content (speak :content "It is in a building in Boston."))
(action	:when (after :event S_BLDG.WORD4) :content (rightgesture :cmd compose :trajectory verticalup :handshape bend))
(action	:when (after :event S_BLDG.END) :content (rightgesture :cmd relax))

Animator

- Drives an articulated character
- Character model read as VRML
- Joints named according to H-Anim spec
- Interpolators animate groups of joints
- Groups are bodyparts like an arm or a hand
- Each group can be given a Shape
- Shapes are used as keyframes

Animator

- Shapes are either predefined or from IK
  - ARM->MoveTo(x,y,z,t_approach,t_duration);
    HAND->SetShape("fist",t_approach,t_duration);
- Series of Shapes define a Path
  - ARM->SetPath("wave",t_approach,t_duration);
    HAND->SetShape("flat", t_approach,t_duration);

Perlinizer: Background Motion Generator

- Adds "lifelike" motion to character
  - Motion while idle
  - Variability during commanded motion
- Each Joint in H-Anim model can have two background signals specified
  - Idle and InUse with smooth transitions between
  - Signals are specified per Perlin Improv system
  - Background signal added to commanded motion

Excerpt from REA

Conversational Functions Currently Modeled

- Notice user presence / absence (gaze)
- Ritual greet and farewell
- Turn-taking
  - Wanting turn interruption
  - Taking turn interruption
- Backchannel feedback
- Simple speech repair
- Role of gesture in semantics & discourse

Conversational Behaviors Exhibited

- Synthesized speech

- Eye gaze towards & away from interlocutor
- Raising eyebrows
- Nods & head shakes
- Many kinds of hand gestures
- Body posture & orientation

Other Important Dimensions (that I haven't covered)

- Individuality
- Artistic appearance
- Learning
- Task specificity / function

- . . .

Conclusion

- Social-conversational skills allow humans to engage one another in information exchange, and the construction of relationships.
- Social-conversational intelligence is a key way of making animated agents more engaging, more credible, more like partners.
- Visual information about conversation plays a key role in manifesting this kind of intelligence.
- Embodied conversational agents may one day allow a face-to-face Turing test . . .

More Information

- http://www.media.mit.edu/~justine/