Published in:
P. McKevitt, S. Ó. Nulláin, C. Mulvihill (eds.), *Language, Vision & Music*, 97-115. Amsterdam: John Benjamins.

MACHINE PERCEPTION OF REAL-TIME MULTIMODAL NATURAL DIALOGUE

KRISTINN R. THÓRISSON

The Media Laboratory, Massachusetts Institute of Technology[1]
Cambridge, Massachusetts 02139, USA
http://xenia.media.mit.edu/~kris kris@media.mit.edu

Keywords: Multimodal perception, natural dialogue, face-to-face communication, software architecture

1. Introduction

A machine capable of taking the place of a person in face-to-face dialogue needs a rich flow of sensory data. Moreover, its perceptual mechanisms need to support interpretations of real-world events that can result in real-time action of the type that people produce effortlessly when interacting via speech and multiple modes. The goal of the work described here has been to create such a machine.

Most of human actions are goal-oriented (Waltz 1999) — avoid hitting that light post, reach for the pen to write, look to see who entered. It is a fair assumption that dialogue is no different from other human activity in this respect. If we couldn't move, act, or communicate, there would be no reason to perceive; perception is the servant of action. In dialogue perception serves the goal of keeping conversants on track, to match the overall plan of the communication. Endowing an artificial agent with multimodal perception gives it a solid foundation to base its actions on (Aloimonos 1993). The approach taken here to perception — and indeed to the general problem of multimodal interpretation — is that it based on highly opportunistic processes, using whatever cues possible to determine the meaning of a communicative act. This view has been voiced by others in the context of natural language interpretation (Pols 1997, Cullingford 1986).

This paper presents the perceptual mechanisms of a computational model of psychosocial dialogue skills called Ymir[2]. Ymir encompasses a layered, modular perception system that allows any embodied, computer-controlled humanoid to participate in natural, multimodal communication with a human. It proposes new ways for achieving real-time performance in language-capable systems. Ymir is constructed with a holistic view on the dialogue process, modeling human-human conversation as a single dialogue system (Thórisson 1999, 1996, 1995). The architecture is also holistic in the sense that it takes into account all types, and time scales, of multimodal behavior perceived and produced in a typical free-form dialogue, and distinguishes itself from a lot of other related research on these grounds (e.g. Heinz & Grobel 1997, Sparrell & Koons 1994, Wahlster 1991). The work described here also distinguishes itself from that of others in that real-time turn-taking is modeled and implemented for multiple modes (cf. Heinz & Grobel 1997, Frölich & Wachsmut 1997, Rigoll et al. 1997, Wren et al. 1997, Essa 1994, Sparrell & Koons 1994, Bolt 1980), and that high-level knowledge and natural language is included in the interaction (cf. Blumberg & Galyean 1995, Brooks & Stein 1993, Wilson 1991, Maes 1990).

Gandalf, the first character created in this architecture, is capable of fluid turn-taking and unscripted, task-oriented dialogue (Thórisson 1999, 1996); he perceives natural language, natural prosody and free-form gesture and body language, and conducts natural, multimodal dialogue with a human. Computer-naïve users have rated him highly on believability, language ability and interaction smoothness (Thórisson 1998, 1996).

Here we will focus on the mechanisms for multimodal perception prescribed by Ymir and realized in Gandalf, emphasizing spatial representation and prosody analysis, and explain how top-down and bottom-up perception is organized in the Ymir architecture. Other aspects of Ymir are discussed elsewhere: Real-time decision making in (Thórisson 1998); real-time motor control in (Thórisson 1997); an overview of Ymir can be found in (Thórisson 1999).

The paper is organized a s follows: First we give an overview of Ymir, with an emphasis on its general perceptual mechanisms. Then we give a summary of Gandalf and the dialogue skills exhibited by this agent. The final section explains how these perceptual mechanisms are implemented in the Gandalf prototype using particular examples. Related work is referenced throughout the paper.

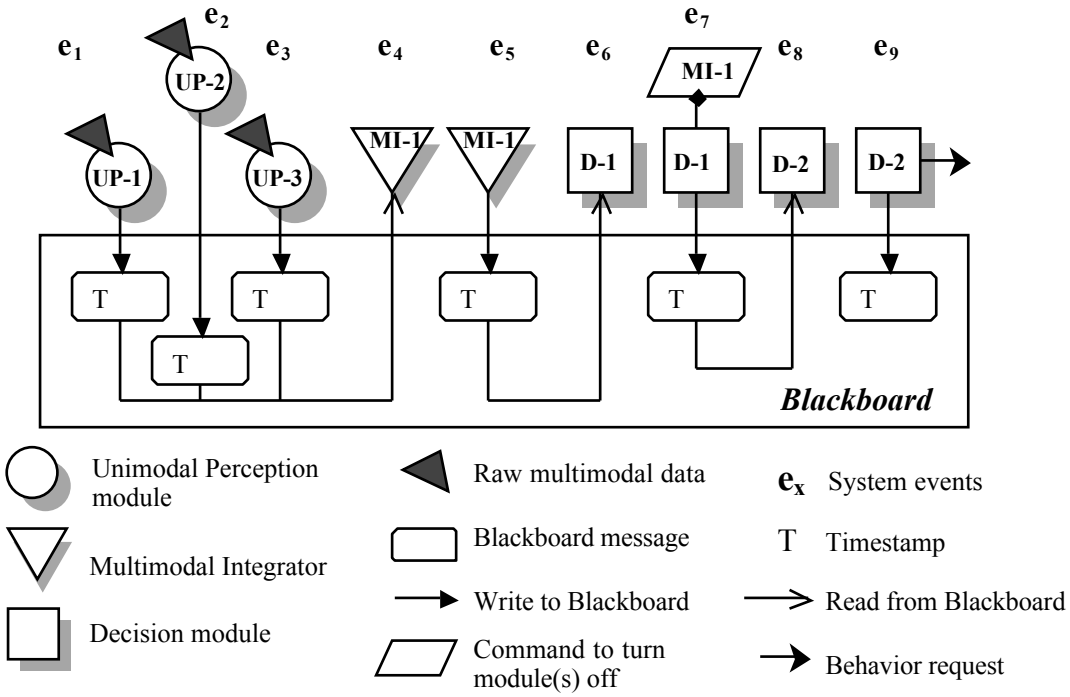

Figure 1. Internal events involving perceptual processing and decision making. Events e1, e2, e3: Raw data from sensing devices streams in to dedicated Unimodal Processors (UP), which process each mode separately and post the results of this processing on one of two blackboards (CB and FSB in Figure 3). e4: Multimodal Integrators (MI) combine the data from two or more UPs and/or MIs and process further, again posting results (e5). e6, e8: Deciders read data from UPs, MIs and other deciders to issue overt and covert decisions. e7: An overt decision is made to turn a covert decider module off, and this decision is posted. e9: A decision is made to issue an overt behavior request.

2. Perceptual Mechanisms in Ymir

Ymir is a computational model of psychosocial dialogue skills that addresses the main characteristics of face-to-face dialogue. It is also a very malleable structure to test models of various mental mechanisms. In Ymir the perceptual abilities of an agent are assumed to be grounded in knowledge about the interaction — knowledge about participants, body parts, turn taking, etc. — via situational indexing. This knowledge is provided through symbolic and spatial representations. Ymir also borrows several features from prior blackboard (Dodhiawala 1989) and behavior-based artificial intelligence architectures (Maes 1990, Wilson 1991), but goes beyond these in the amount of communication modalities and performance criteria it addresses.

Ymir has four types of processing modules: *perceptual*, *decision*, *knowledge* and *behavior-motor*, in four process collections, {1} a *Reactive layer* (RL), {2} a *Process Control layer* (PCL), {3} a *Content layer* (CL), and an {4} *Action Scheduler* (AS). The four process collections give Ymir a hierarchical structure for dealing with *time* and *complexity*. Multimodal user behavior is collected through chosen tracking mechanisms and separated into significant segments (e.g. *hands*, *arm*, *trunk* for the body; *intonation* and *words* for speech; see below). This data streams into all three layers (Figure 3), which contain Perception and Decision modules. The output of the Perception modules provides input to Decision modules (see Figure 1). Perception modules with particular cycle times process the raw data to various degrees and output their results to one of two blackboards, supporting decisions with corresponding perceive-act cycle times (Thórisson 1998). Interpretation can be driven, or triggered, based on any perceptual data — speech, prosody, gesture, gaze, internally generated decisions and knowledge, or any combination of these. These principles serve as the foundation for perceptual data handling and integration, and support real-time decision making, planning, and interpretation throughout the architecture.

The relationship between Perception modules and Decision modules in Ymir is the relationship between

bottom-up and top-down processing: (1) Bottom-up processing works by having Perception and Decision modules in one layer process *only data from modules in the same layer or the layer below*. Economical use of computation is thus gained through bottom-up "value-added" data flow, modules in each successive layer add information to the results from the layer below. (2) Top-down control is achieved by having Perceptual modules in one layer *turned on and off by Decision modules in the layer above and sometimes in the same layer* (but never below). A goal produced by planning algorithms in the Content Layer (CL) can trigger a Decision module to turn on or off a Perception module (or a group of them) in the Process Control Layer. Interweaving top-down and bottom-up processing in this way is powerful enough for a broad range of dialogue mechanisms, and is easy to manage.

```
MODULE-TYPE: dir-Unimodal-Perceptor
NAME: user-looking-at-me?
DATA-1: user-gaze-direction-vector
DATA-2: my-head-position
INDEX-1: get-users-gaze-direction
INDEX-2: get-my-head-position
FUNCTION: user-looking-at-me?
BLACKBOARD-DEST: FSB
TIMESTAMP: 203948
STATE: TRUE
```

Figure 2. Example of an implemented spatio-directional Unimodal Perceptor. The process *user-looking-at-me?* computes the intersection of a cone and a plane using spatial representations (see Figures 5 and 6). The state of the module changes based on whether its function returns true or false.

Perceptual Modules and Perceptual Integration

The two main types of Perception modules are Unimodal Perceptors (UP) and Multimodal Integrators (MI). UPs process data from only a single mode or a subset of a mode (e.g. speech, manual gesture, prosody); these are relatively simple processes (Figure 2). A total of 26 Perception Modules were created for Gandalf; 16 Unimodal Perceptors (4 prosody, 3 speech, 9 body), and 10 Multimodal Integrators describing the user's turn-giving, manual gesture activity, back-channel activity and completeness (syntactic, grammatical, and pragmatic) of the utterance/multimodal act. Unimodal Perceptor modules are classified into groups based on their function. These have been implemented in the Gandalf prototype: (1) *spatio-positional*, (2) *spatio-directional*, (3) *speech*, (4) *prosody*. The MIs aggregate information from the UPs, and other MIs, to compute multimodal descriptions of the user's behavior. An example is a collection of MIs that determine in concert whether the user is giving the turn. Perceptual modules can also be *static* or *dynamic*. Static modules look at a state in time, regardless of history; dynamic modules integrate data over time.

The use of time-stamped symbolic tokens (e.g. `{User-Speaking, TRUE, 9394}, {Intonation-Going UP, 9672}` as the main format of the perceptual system output proved to be very useful. It is easy to work with and is directly digestible by automatic decision processes using first-order or fuzzy logic — two very appealing candidates for use in decision mechanisms. Significant results in the classification of cue phrases have been achieved by Litman (1996) using similar logical combinations of features gleaned from natural language and speech.

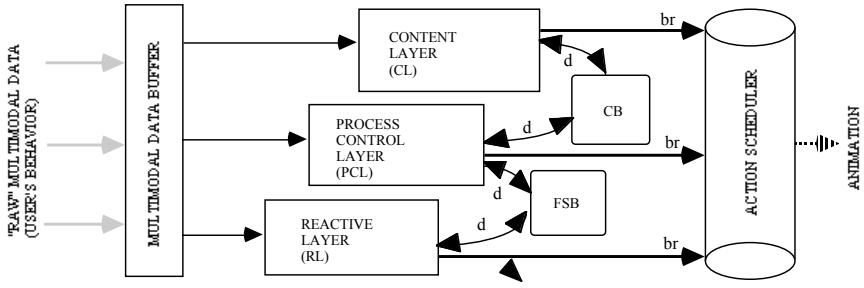

Figure 3. The Ymir architecture is composed of four process collections, each which is made up of several modules for perception, knowledge, decision and action. Multimodal data streams in from the left and is stored in a buffer. From this buffer perception, decision and knowledge modules fetch the data they need to produce their output. This output (d) is posted in blackboards (Content Blackboard and Functional Sketchboard). When a decision is made a behavior request (br) is sent to the Action Scheduler, and usually an animation action results, controlling one or more muscles of the agent's body. Not depicted is a Motor Feedback blackboard, used for tracking motor state.

Ymir presents an inherent symmetry between perception, decision, and action, inspired by behavior-based AI (cf. Wilson 1991). For example, the representation and distribution of Perceptual modules in the layers are complementary to the Decision modules: In the RL both have a relatively low accuracy/speed trade-off; the PCL contains more sophisticated perceptual processing — and more sophisticated decision making; and so on in the CL. The Action Scheduler mirrors this

symmetry: The layer from which an incoming behavior requests comes determines its priority, reactive decisions having the highest. An action such as turning the head toward a sudden sound is thus always guaranteed the shortest and most appropriate route through the system, from the perceptual cue, to the decision to turn, to the process that pulls the muscles (Thórisson 1998).

Gandalf: Summary of Capabilities

To construct Gandalf's dialogue skills, data from the psychological and linguistic literature was modeled in Ymir's layered, modular structure (cf. McNeill 1992, Rimé & Schiaratura 1991, Clark & Brennan 1990, Pierrehumbert & Hirschberg 1990, Groz & Sidner 1986, Kleinke 1986, Goodwin 1981, Duncan 1989, Sacks et al. 1974, Yngve 1970, Effron 1941). This enables Gandalf to participate in *collaborative, task-related activities* with users, perceiving and manipulating a three-dimensional graphical model of the solar system: Via natural dialogue a user can ask Gandalf can manipulate the model, travel to any of the planets, and tell about them (Figure 4).

Gandalf's perception extracts the following kinds information from a conversational partner's behavior: (1) Eyes*: Attentional and deictic functions* of conversational partner, during speaking and listening. (2) Hands*: Deictic gesture* — pointing at objects, and *iconic gesture* illustrating tilting (in the context of 3-D viewpoint shifts). (3) Vocal*: Prosody* — timing of partner's speech-related sounds, and intonation, as well as ***speech content*** — words produced by a speech recognizer. (4) Body: ***Direction of head and trunk*** and *position of hands in body space*. (5) *Turn-taking signals:* Various feature-based analyses of co-dependent and/or co-occurring multimodal events, such as intonation, hand position and head direction, and combinations thereof. These perceptions (1-5) are interpreted in context to conduct a real-time dialogue. For example, when Gandalf takes turn he may move is eyebrows up and down quickly (a common turn-taking signal in the western world) in the same way as humans (typically 2-300 msec after user gives turn). This kind of precision is made possible by making timing a core feature of the architecture. Gandalf will also look (with a glance or by turning his head) in the direction that a user points. The perception of such gestures is based on data from multiple modes; where the user is looking, shape of the user's hand, and data from intonation. The result is that Gandalf's behavior is highly relevant to the user's actions, even under high variability and individual differences.

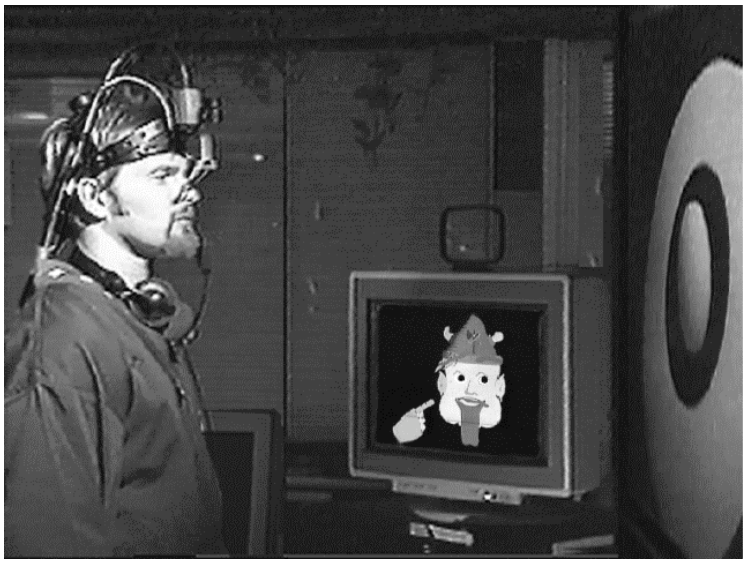

Figure 4. Gandalf appears on its own monitor, the large monitor (right) displays a model of the solar system. Here Gandalf (viking helmet) points (using a manual gesture) as he answers the author's (eye tracking helmet) question "What planet is that?" with the utterance "That is a top view of Saturn".

Gandalf uses these perceptual data and interpretations to produce real-time multimodal behavior output in the all of the above categories, with the <u>*addition of:*</u> (6) Hands: *Emblematic gesture* — e.g. holding the hand up, palm forward, signaling "hold it" to interrupt when the user is speaking, and *beat gestures* — hand motion synchronized with speech production. (7) Face: *Emotional emblems* — smiling, looking puzzled, and *communicative signals* — e.g. raising eyebrows when greeting, blinking differently depending on the pace of the dialogue, facing user when answering questions, and more. (8) Body: ***Emblematic body language*** — nodding, shaking head. (9) Speech:[3] *Back channel feedback*. These are all inserted into the dialogue by Gandalf in a free-form way, to support and sustain the dialogue in real-time. While the perception and action processes are highly time-sensitive and opportunistic, the Ymir architecture allows Gandalf to produce completely coherent behavior.

4. Implementation

We will now look at the implementation-specific details of Gandalf's visual, speech, and prosody perception.

In any efficient perception system the world is sampled at a rate of 20-30 Hz. In the Gandalf prototype an eye tracker, body tracking suit and gloves[4] produce over 60 floating point numbers at this rate (Bers 1996); the prosody analyzer reports (filtered) changes in intonation *during* speech at 3-4 Hz, and the speech recognizer delivers words *after* speech. As mentioned before, Ymir deals with timing explicitly — all perceptual events are time-stamped as they are received through transducers, and every time they are re-processed and re-posted to a blackboard by a module. Time-stamps for each successive improvement in a perception's detail and/or accuracy are carried through, such that the history of raw data processing towards full interpretation can be traced back to its origins at any point in the processing, from perception of raw data to a potential action resulting from it.

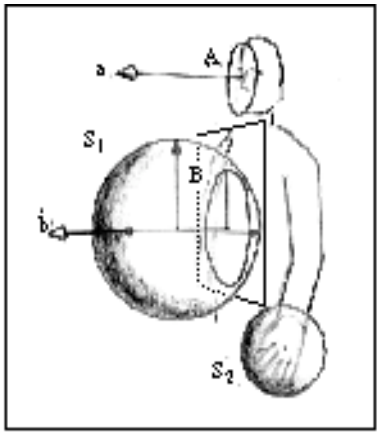

Figure 5. Geometric definitions of gesture volume (S1), face plane (A) and hand volume (S2), along with normals showing direction of head (a) and trunk (b). Plane B is an offset of S1 from the trunk.

Vision in Situated Dialogue

All visual perception in the Gandalf prototype happens through geometric representations of the world, an approach shared by others (Wren et al. 1997).[5] Gandalf's real-world environment — the user, position of screens — is mapped out using the same geometric technique. A large, flat-screen monitor is used to display a 3-D model of the solar system; planets appearing on this monitor were mapped to the screen's 2-D projection in real space, so that the agent could point its gaze and hand at them. High-frequency, high-reliability measurement of body movements allows high-frequency user behavior such as fixations to be tracked and integrated in real-time in a realistic manner. An example of this is Gandalf glancing to the big screen, mirroring a user's gaze towards an object, in less than 300 ms (Figure 4). When combined with real-time prosody tracking, the analysis of the geometric representation described was sufficient to support the realistic production of reactive, interactive dialogue behaviors such as fluid turn-taking (Thórisson 2001), back channel feedback (Yngve 1970) and gaze control (Kleinke 1986), behaviors which require perception-action loops of 250 ms or less.

Information Spaces

Space is divided into *volumes*, *planes* and *points* for perceptual processing. Three spatial features are critical to conversants in multimodal dialogue: {1} *Gesture volumes*, {2} *Face planes*, {3} *Work volume(s)*. These features mark the (somewhat fuzzy) boundary in which events, objects or sources of information can be located during interaction. Knowing the sizes and shapes of these is not enough, however, one needs to know the positions of these volumes and planes, and, in particular, objects within them. The following coarse positional data are essential to embodied dialogue: {1} *Position of work volume*, {2} *Position of gesture volumes*, {3} *Position of face planes*, and {4} *Position of hands* (or hand volumes).

In the Gandalf prototype volumes are mapped out as shown in Figure 5. Work space and faces are simply three-dimensional planes with an orientation — a circular one for the face and a square one for the work space display (not shown). Position is given by the planes' centers. In the prototype, objects within these volumes and planes (other than hand and face) can be treated as 2-D and 3-D points — a useful (but not always completely accurate) simplification. The user's hands are surrounded by a 20 cm diameter sphere, in order to give their position a larger margin. Coarse body movements can be generated on the basis of coarse spatial knowledge — e.g. the boundaries of a volume. For example, orienting your body towards a person requires only rough knowledge of where the person is located relative to you. Fine motion, such as gaze, requires very accurate pinpointing of objects in space.

Directional Data

When is a person "facing" someone? When is a person "turned to" something? These questions need to be answered by participants in any embodied, face-to-face conversation, if they want the interaction to succeed. The directional features extracted in Gandalf are: {1} *Direction of gaze*, {2} *Direction of head*, {3} *Direction of trunk*, {4} *Direction of deictic gestures*. Figure 5 shows the two normals used for head and trunk. The head, gaze and trunk normals are made into cones so that their interception with other spaces, such as the agent's face space, is broader: Interception happens if the inside of the cone overlaps the area or point in question (Figure 6).

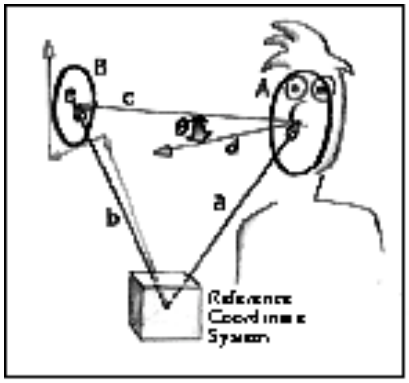

Figure 6. Geometry defining the "facing" function. Center of Face Plane A is defined by vector a; d is plane A's normal. Center of Plane B is defined by vector b; □ defines plane A's cone. By comparing the angle between vectors d and c to a threshold, one can determine whether the person on the right is facing plane B, which could e.g. represent the agent's face. This formulation was designed to correspond to our intuitive notion of "facing".

The angle of the cones is graded such that gaze has the narrowest (a 20° cone), then the head (35°), and lastly the trunk (40°). These were chosen based on the frequency of movements of these body parts to the amount of error in measuring them, and the size of the objects of interest (Gandalf's face for example). The user is *looking at point p* if *p* falls within the boundary of the gaze cone; the user is *facing point p* if *p* falls within the boundary of the head cone (Figure 6); and the *user is turned to point p* if *p* falls within the boundary of the body cone. Saccades provide an excellent basis for gaze segmentation: We filter eye movements into *saccades*, *fixations* and *blinks*, and use only data from the fixation when estimating the gaze vector (Koons & Thórisson 1993).

Gesture Recognition

Ymir goes beyond most recent efforts in co-verbal gesture recognition, which typically are limited to only one of the gesture types, do not allow their free mixture, do not consider other contextual factors like gaze, speech and dialogue state, and/or do not consider real-time production of reciprocal multimodal acts (cf. Frölich & Wachsmut 1997, Hoffman et al. 1997, Rigoll et al. 1997, Horpraset et al. 1996, Sparrell & Koons 1994, Wahlster 1991). By recognizing two classes of manual gesture in unconstrained dialogue, Gandalf demonstrates the generality of the Ymir architecture for integrating a wide variety of multimodal events in real-time by contextually driven processing. Future extensions to Ymir include the remaining classes of gestures — *pantomimic*, *self-adjustors*, *metaphoric*, *emblematic* and *beat* (McNeill 1992, Effron 1941) — involve the addition of modules, rules and interpretation mechanisms, but not new architectural constructs.

In contrast to automatic sign-language recognition (cf. Hermann & Grobel 1997) — where processing ends with deducing the correct meaning of well-formed gestures — free-form, co-temporal, co-spatial dialogue additionally means producing real-time gesture and other behavior interactively. The gesture recognition method used in the Gandalf prototype derives from Sparrell and Koons' work on co-verbal gesture (1994), modified to fit the larger psychosocial context addressed in Ymir, mainly by linking the interpretation mechanisms to the bottom-up, top-down processing scheme. Hand posture is classified into different shape categories by Unimodal Perceptors; hand placement is categorized as *in* and *out* of gesture space, and the hands' distance from the trunk is also tracked. In sign language, placement of hands has been given a-priori meaning (Heinz & Grobel 1997) — natural gesture is more complex, and most cannot be interpreted without reference to context and cross-modal indexing (McNeill 1992, Goodwin 1981). Multimodal Integrators combine this data with data from the perceived dialogue state (e.g. "does the user have the turn?"), speech activity (e.g. "is the user speaking?"), shape of the hand and its placement relative to the user's body, to produce a perception of a *communicative gesture*.

When processing data from the bottom-up, with no evidence from other modes or context, gesture space is a very good first approximation to isolating communicative gesture. Self-adjusters (e.g. scratching, adjusting clothing), which seldom have a communicative function, are mostly excluded by lifting the gesture space 10 cm from the body of the user (plane B in Figure 5). Should they happen outside of gesture space the gesture could still be caught based on context and the user's speech acts. (However, its processing might not be as fast since the analysis would rely on slower top-down methods.) Gesture space thus serves several roles in the Ymir architecture. {1} It is a strong real-time indicator of the presence of *intentional, communicative* gesturing, which mainly happens in the space right in front of the speaker's body (Rimé & Schiaratura 1991). {2} It is useful for estimating if people are looking at their hands, often an indication of iconic gestures. {3} It is useful in data stream segmentation, when there is uncertainly about where (in time) a gesture occurred, and {4} gesture space has turned out to be useful for directing the agent's attention, for example to look at the speaker's face or to gestural referents at the right moments in time. McNeill's research (1992) has indicated that the type of gesture and its place of articulation may be correlated. If this turns out to be the case, a finer division of gesture space would be useful for determining the function of manual gestures (but not their presence).

Iconic and Deictic Gestures

Among the features used to recognize deictic gesture where *hand posture* (index finger extended, other fingers mostly bent), *position of hand relative to the gesturer's body* ("in gesture space") as well as *vocalization in a given temporal proximity of the hand-arm motion*.[6] As mentioned, all perception events are time-stamped:

Data about the direction of the index finger (and/or arm) at the time of the pointing is used to compute the direction the user pointed in, to produce a referent object, often after the fact. To be able to glance in the direction of the user's pointing *while* the user is pointing, Gandalf can sample the user's gaze direction, which is continuously computed. This often proves satisfactory for producing correct glancing. Of course, a precursor to actually looking in the direction pointed is that we know that the hand/arm movement represents a communicative gesture, and that the type of the gesture is in fact deictic (the only communicative gestures whose single function is to direct spatial attention). This information comes from Multimodal Integrators looking at a holistic picture of the user's actions: body posture, gaze, prosody, hand posture, etc., which are provided by processes mainly in the Reactive Layer and Process Control Layer (see Figure 3).

In addition to co-verbal deictic gestures, Gandalf recognizes iconic gesture (e.g. "Tilt it like this [wrist motion indicating direction]"). Iconic gestures have less strict a morphology[7] than deictic ones, and are harder to detect in real-time (at the actual time of occurrence). Like deictic recognition, iconic recognition relies heavily on Multimodal Integrator classification between *communicative* versus *non-communicative* gestures. Without this high-level distinction reliability of iconic recognition falls significantly.

In summary, four main features characterize the vision work described above: {1} The vision mechanisms support interpretation of behavior during completely natural, spontaneous dialogue; {2} The vision system support contextualized, high-level percepts, combining knowledge-based (top-down) and data-driven (bottom-up) processing; {3} Vision mechanisms are embedded in a real-time architecture, constrained by a requirement to perform dialogue behavior in a natural manner, based on data from natural human interaction (c.f. Goodwin 1981, Duncan 1989, Yngve 1970); and {4} The vision mechanisms are integrated with other perception (prosody and speech content) in a unified way. Together these characteristics set this work apart from a majority of other computer vision research.

Hearing: Word Recognition

Gandalf hears *speech* sounds, recognized via two mechanisms: A *prosody analyzer* and an off-the-shelf *speech recognizer*. Words are teased out of the speech stream through continuous-speech, grammar-based, methods. Instead of waiting for a significant pause at the end of an utterance, as is frequently done in commercial speech recognizers, the prototype demonstrated the superior method of triggering speech recognition based on multimodal data and turn-state. To implement this in Ymir a Decider is constructed that looks at the output of selected Multimodal Integrators and initiates, given the right conditions, recognition to be done on the audio collected since *last turn*. The same goes for interpretation: Once words have been received in the Content Layer interpretation is initiated by Deciders that monitor the state of the speech recognizer and the turn-taking system. The delay time for the recognition of an average sentence (form the end of the utterance) is around two seconds. In a speech-only system this would be a significant problem, since turn-taking in human telephone dialogue expects faster responses to *speech content* (~500-1000 ms), and immediate responses to *turn-taking cues* (~0-250 ms) (Goodwin 1981). This is different for face-to-face dialogue, where other modes come into play. In spite of this "mental limitation" in processing speech content, Gandalf's turn-taking is socially acceptable (frequently within 300 msec — and with appropriate repair of failures) because of data that the real-time prosody analyzer and body tracking provided, which allows it more intelligent management of dialogue behavior during turn transitions via relevant gaze, eyebrows and head movements.

Hearing: Real-Time Prosody Analysis

Methods have been suggested for automatic analysis of prosody (Todd & Brown 1994, Pierrehumbert & Hirschberg 1990), but few have focused on real-time analysis (Nöth et al. 1997). As mentioned in the introduction, we view the human perceptual system as highly opportunistic and flexible: Any cue can be used to aid interpretation and creation of meaning structures.

The analyzer developed here detects the following boolean, time-stamped events in real-time: {1} *Speech on,*[8] {2} *Intonation going up*, {3} *Intonation going down*, {4} *Intonation flat*. The intonation analysis is performed using a windowing technique, where a window is 300 ms. A new window starts where the last one ended. The slope of the intonational contour is estimated for each window and checked against a threshold. If over the threshold and different from last window, a time-stamped status report is posted on a blackboard about intonational direction. This is used for increasing the reliability of end-of-utterance detection, and inferring the type of utterance (e.g. question or command). In a future version of the system interpretation mechanisms will use this for identifying where the emphasis lies in a sentence.

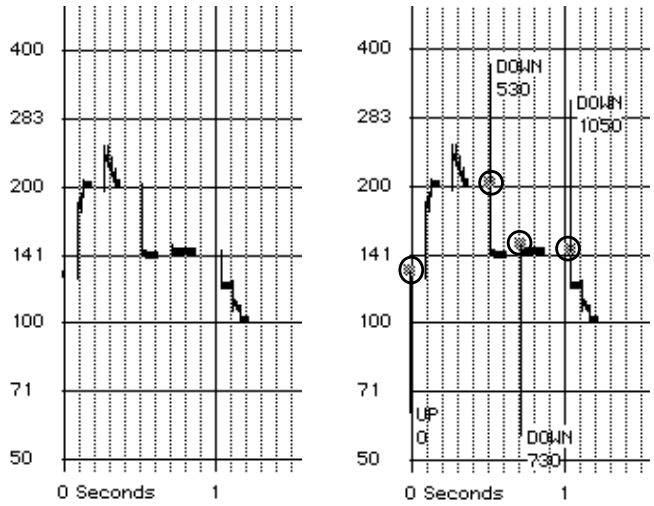

Figure 7. Example of intonation for the utterance "Take me to Jupiter" plotted to a logarithmic frequency scale (y-axis). On the right we see the result of the real-time intonation analysis. Segmentation of pitch direction is marked with vertical bars, giving timing (in ms) and direction of the audio stream. All features in this example took less than 10 ms to compute. Any delay in computing is estimated based on the raw data and then subtracted when time-stamping the event.

By running this prosody system on a dedicated machine, relatively robust, real-time performance is achieved (Figure 7). The reliability is high enough for an interactive system (roughly one utterance in ten is impossible to analyze in real-time) — other modes help correct for occasional failures in this unimodal analysis.

5. Conclusions & Future Work

The perception scheme modeled in Ymir results in fluid dialogue, as evidenced by interactions between novice users and Gandalf (detailed performance results are reported in Thórisson (1999) and (1996)). Some of that success can be attributed to the data collection (dressing the user up in the agent's perceptual "organs"), which produces in high-speed, partially processed, and relatively noise-free data. But the bulk of it is undoubtedly due to the hierarchical perception-action structure of Ymir, which was modeled as a unified system and designed from the ground up to support real-time multimodal behavior.

Ymir already provides the framework for perceiving facial gesture in context, and the full range of manual gesture. Including these is a matter of adding input devices such as cameras for face sensing, and a complementary lexico-geometric knowledge bases of face and gesture. The number of perception modules (26 in total) in Gandalf could easily extended to support these, or be extended simply for a significantly more capable agent, increasing robustness, number of manual gesture types recognized, adding facial gesture recognition, and more.

Although intended primarily for embodied, multimodal, task-oriented dialogue, Ymir is not restricted to perceptions of communication. The architecture's perceptual mechanisms seem for example nicely suited for characters inhabiting virtual worlds. Preliminary testing indicates that skills such as navigation, story telling, walking, flying, etc. in virtual worlds can easily be implemented in the same framework (Bryson & Thórisson 2001).

Future work on these perception mechanisms focuses on more sophisticated attentional control and various methods for perceptual classification. Part of this work will involve strengthening the connection between perception and knowledge, thus improving the agent's responses and understanding of the dialogue. Integrating increasingly sophisticated knowledge representation into the architecture will be a significant part of this task.

Acknowledgments

I would like to thank the sponsors of this work: Thomson-CSF, the M.I.T. Media Laboratory, RANNÍS, TSG Magic and the HUMANOID Group. My thanks also to Pattie Maes, Richard A. Bolt, Justine Cassell and Steve Whittaker for inspiration and brilliant guidance; Joshua Bers, David Berger & Hannes Vilhjálmsson for programming, advice, and suggestions.

References

Aloimonos, Y. (ed.). (1993). *Active Perception*. Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.

Bers, J. A (1996). Body Model Server for Human Motion Capture and Representation. *Presence: Teleoperators and Virtual Environments*, 5(4), 381-392.

Blumberg, B. M. & Galyean, T. A. (1995). Multi-Level Direction of Autonomous Creatures for Real-Time Virtual Environments. Proceedings of SIGGRAPH '95 August, 47-54.

Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. *Computer Graphics*, 14(3), 262-70.

Brooks, R. & Stein, L. A. (1993). Building Brains for Bodies. M.I.T. Artifical Intelligence Laboratory memo No. 1439, August.

Bryson, J. & Thórisson, K. R. (2001). Dragons, Bats & Evil Knights: A Three-Layer Design Approach to Character-Based Creative Play. To be published in D. Ballin (Ed.), *Virtual Reality, Special Issue on Intelligent Virtual Agents*, spring. Heidelberg: Springer-Verlag.

Clark, H. H. & Brennan, S. E. (1990). Grounding in Communication. In L. B. Resnick, J. Levine & S. D. Bahrend (eds.), *Perspectives on Socially Shared Cognition*, 127-149. American Psychological Association.

Cullingford, R. E. (1986). *Natural Language Processing: A Knowledge-Engineering Approach*. NJ: Rowman & Littlefield.

Dodhiawala, R. T. (1989). Blackboard Systems in Real-Time Problem Solving. In Jagannathan, V., Dodhiawala, R. & Baum, L. S. (Eds.), *Blackboard Architectures and Applications,* 181-191. Boston, MA: Academic Press, Inc.

Duncan, S. Jr. (1989). Some Signals and Rules for Taking Speaking Turns in Conversations. Journal of Personality and Social Psychology, 23(2), 283-292.

Effron, D. (1941/1972). *Gesture, Race and Culture*. The Hague: Mouton.

Essa, I. A., Darrell, T. & Pentland, A. (1994). Modeling and Interactive Animation of Facial Expression using Vision. M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 256.

Frölich, M. & Wachsmut, I. (1997). Gesture Recognition of the Upper Limbs — From Signal to Symbol. In I. Wachsmut & M. Frölich (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, 173-184. Berlin: Springer.

Goodwin, C., (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Horpraset, T., Haritaoglu, I., Harwood, D., Davis, L., Wren, C. & Pentland, A. (1996). Realt-Time 3D Motion Capture. *Proc. of the 2nd Workshop of Perceptual User Interfaces*, Nov. 4-6, San Francisco, U.S.A.

Grosz, B. J. & Sidner, C. L. (1986). Attention, Intentions, and the Strucutre of Discourse. *Computational Linguistics*, 12(3), 175-204.

Hienz, H. & Grobel, K. (1997). Automatic Estimation of Body Regions from Video Images. In I. Wachsmut & M. Frölich (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, 135-145. Berlin: Springer.

Hoffman, F. G., Heyer, P., Hommel, G. (1997). Velocity Profile Based Recognition of Dynamic Gesture with Discrete Hidden Markov Models. In I. Wachsmut & M. Frölich (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, 81-95. Berlin: Springer.

Kleinke, C. (1986). Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, 100(1), 78-100.

Koons, D. B. & Thórisson, K. R. (1993). Estimating Direction of Gaze in Multi-Modal Context. Presented at *3CYBERCONF—The Third International Conference on Cyberspace*, Austin TX, May 13-14.

Litman, D. J. (1996). Cue Phrase Classification Using Machine Learning. *Journal of Artificial Intelligence Research*, 5, 53-94.

Maes, P. (1990). Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back. In P. Maes (Ed.), *Designing Autonomous Agents*, 1-2. Cambridge, MA: MIT Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Nöth, E., A. Batliner, A. Kiessling, R. Kompe & H. Niemann (1997). Suprasegmental modelling. *Informal Proceedings of NATO ASI on Computational Models of Speech Pattern Processing*, St. Helier, Jersey Channel Islands.

Pierrehumbert, J. & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In P. R. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions in Communication*. Cambridge: MIT Press.

Pols, L.C.W. (1997). Flexible, Robust, and Efficient Human Speech Recognition. Technical Report, Institute of Phonetic Sciences, University of Amsterdam Proceedings 21, 1-10.

Rigoll, G., Kosmala, A. & Eickeler, S. (1997). High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In I. Wachsmut & M. Frölich (Eds.), *Gesture and Sign Language in Human-Computer Interaction,* 69-80. Berlin: Springer.

Rimé, B. & Schiaratura, L. (1991). Gesture and Speech. In R. S. Feldman & B. Rimé. *Fundamentals of Nonverbal Behavior*, 239-281. New York: Press Syndicate of the University of Cambridge.

Sacks, H., Schegloff, E. A. & Jefferson, G. A. (1974). A Simplest Systematics for the Organization of Turn-Taking in Conversation. *Language*, 50, 696-735.

Sparrell, C. J. & Koons, D. B. (1994). Capturing and Interpreting Coverbal Depictive Gestures. *AAAI 1994 Spring Symposium Series*, Stanford, USA, March 21-23, 8-12.

Thórisson, K. R. (2001). Natural Communication Needs No Manual: A Computational Model of Real-Time Turn-Taking in Natural Multimodal Dialogue. To be published in B. Granström (Ed.), *Multimodality in Language and Speech Systems*. Heidelberg: Springer-Verlag.

Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures and Humanoids. *International Journal of Applied Artificial Intelligence*, 13 (4-5), 449-486.

Thórisson, K. R. (1998). Real-time Decision Making in Face-to-Face Communication. *The Second ACM Conference on Autonomous Agents*, Minneapolis, MN, 16-23.

Thórisson, K. R. (1997). Layered Modular Action Control for Communicative Humanoids. *Proceedings of Computer Graphics Europe*, June 5-7, Genieva, 134-143.

Thórisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. Thesis, Massachusetts Institute of Technology.

Thórisson, K. R. (1995). Computational Characteristics of Multimodal Dialogue. AAAI *Fall Symposium Series on Embodied Language and Action*, November 10-12, Massachusetts Institute of Technology, Cambridge, 102-108.

Todd, N. P. M. & Brown, G. J. (1994). A Computational Model of Prosody Perception. *Proceedings of the International Conference on Spoke Language Processing (ICLSP-94),* Yokohama, Japan, Sept. 18-22, 127-30.

Wahlster, W. (1991). User and Discourse Models for Multimodal Communication. In J. W. Sullivan & S. W. Tyler (eds.), *Intelligent User Interfaces*, 45-67. New York, New York: ACM Press, Addison-Wesley Publishing Company.

Waltz, D. (1999). The Importance of Importance. *AI Magazine*, AAAI-98 Presidential Address, fall, 19-35.

Wilson, S. W. (1991). The Animat Path to AI. In J-A. Meyer & S. W. Wilson (eds.), *From Animals to Animats*. Cambridge, MA: MIT Press.

Wren, C. Sparacino, F., Azarbayejani, A. J., Darrell, T.J., Starner, T.E., Kotani, A., Chao, C.M., Hlavac, M., Russell, K.B., Pentland, A.P. (1997). Perceptive Spaces for Performance & Entertainment: Untethered Interaction using Computer Vision & Audition. *Applied Artificial Intelligence*, June, 11 (4), 267-284.

Yngve, V. H. (1970). On Getting a Word in Edgewise. *Papers from the Sixth Regional Meeting*, Chicago Linguistics Society, 567-78.

---

[1] Now at Soliloquy Inc., 251-255 Park Avenue South, 6th floor, New York, NY 10010, U.S.A.

[2] "Ymir" is pronounced *E*-mirr. The names Ymir and Gandalf are from the Icelandic Sagas.

[3] Gandalf speaks using an off-the-shelf speech synthesizer and uses its built-in rules for intonation and prosody.

[4] All perceptual solutions in the Gandalf implementation are independent of tracking technique — the relatively intrusive tracking methods used for Gandalf could be replaced with computer vision. Pre-processing in the body model used here (Bers 1996) is based on the same knowledge-based techniques used in many computer vision systems (Wren et al. 1997).

[5] By tracking the user's body with a data suit, eye-tracker and data gloves (Bers 1996) a fully geometric picture of the user's upper body is made available to the virtual agent at 20-30 Hz. This tracking is currently more robust than tracking done with state-of-the-art computer vision techniques using cameras (Wren et al. 1997, Horprasert et al. 1996).

[6] This last feature is a useful heuristic for filtering out non-communicative gestures while the agent has the turn.

[7] Morphology is the form of a gesture, i.e. the way its posture or motion *looks*, as opposed to function.

[8] Although detecting a feature such as "speech on/off" may seem trivial, this is only true when using a dedicated microphone which is unlikely to pick up anything besides the dialogue participant's speech. Using signal processing with remote-mounted microphones makes this a significantly more difficult task. In any case, the perception is one that humans have to solve successfully in real-time during dialogue and thus a valid feature with which to provide a virtual humanoid.