19

Towards a neurocognitive model of turn taking in multimodal dialog

James Bonaiuto and Kristinn R Thórisson

19.1 Introduction

Communicating individuals take turns speaking, gesturing, and interacting according to their goals and circumstances. The process is referred to as turn taking and it is a major organizing mechanism in real-time dialog. Thorough psychological studies have shown there to be a significant amount of similarity between societies with regard to observed behaviors during conversation (O'Connell *et al.* 1990). More recently, turn taking has become an issue in robot and virtual human research as researchers aim to make these systems more fluent and dynamic when interacting naturally with humans (cf. Gratch*et al.* 2006; Leβmann *et al.* 2004; Maxwell *et al.* 2001; Lemon *et al.* 2001; Bischoff 2000; Thórisson 1993). In spite of an apparent simplicity (what could be simpler than taking turns in speaking?), turn taking is a complex phenomenon that has eluded clear definition to date, although Sacks *et al.* (1974), Goodwin (1981), and others have certainly made measurable progress in that respect.

Multimodal natural communication involves many perceptual, planning, and motor mechanisms. A comprehensive model of turn taking must address not only how people produce hierarchically organized actions but also how they recognize these when produced by others. One path towards building a biologically plausible, inclusive model of cognitive mechanisms involved in real-time turn taking is by integrating models at different levels of description, for example cognitive and neural. A key assumption in the present work is that turn taking during conversation exists primarily (although certainly not solely[1]) for the purpose of helping participants reduce cognitive load during the interpretation of the content of the conversation. Any production related to the topic of the dialog naturally interferes with the interpretation process; therefore, understanding deteriorates when we try to speak at the same time as we try to listen. Because of this, perception of behaviors that are not related directly to dialog content but rather have the goal of guiding the overall interaction, such as back channel feedback (Yngve 1971) and

[1] Although the *raison d'etre* of turn taking is reduction of cognitive load, numerous features of turn taking can nonetheless serve specific conversational functions, for example when deliberate simultaneous speech is used to display aggression, avoidance of overlapping speech is used to indicate politeness, and mutual silence is interpreted as a wish to change the subject.

turn-taking displays (Duncan 1972), is realized by processes that are significantly more automatic than the processes used to interpret content. In this way, behaviors intended to guide the dialog *process*, as opposed to the topic—what have been called *envelope behaviors* (Thórisson 1996)—can proceed with minimal interference with the semantics of the dialog content.

In any communication where the communicating parties have aligned goals (that is, their goals are matched, at a high level, to produce an outcome favorable to both; e.g. one party wants to understand the movie plot, the other wants to explain it), one can expect to see a minimization of anything that may risk overloading the memory system of either or both parties, which in turn results in smooth turn taking. To a certain extent dialog participants tune their behavior to match that of their interlocutors, based on a (set of) particular purpose(s). In any extended dialog the result is *an alignment of goals*—we can talk about the participants' behavior being *coupled*; in fact, we can take one step further and say that they contain cognitive processes that are coupled.

In our view turn taking is primarily an *emergent phenomenon*—a high-level expression of complex interaction between underlying mechanisms and machinery encompassing plan and goal structures, social conventions, personal attitudes, as well as constraints on the human motor system and the human cognitive apparatus. The most promising way to understand such complex phenomena at present is to model them computationally in dynamic, runnable models, preferably in ways that can be tested in interaction with real humans. Such a view is compatible with Dynamic Syntax approaches to dialog modeling (cf. Cann *et al*. 2005), which view turn taking as the emergent result of incremental parsing and production rather than of elaborate structures such as dialog grammars.[2] The Ymir Turn-Taking Model (YTTM, Thórisson 2002) takes this approach. Based on data from psychological experiments, it is inherently multimodal and incorporates the perception–action loop necessary in real-time embodied turn taking.

We are working to expand selected abstractions in the YTTM, replacing them with detailed neural models. If such an expansion can be done without violating the underlying key assumptions and architectural constructs of both models, the case for both models would be strengthened. Decomposition of modular architectures such as the YTTM facilitates their extension both upwards and downwards, to neurally-plausible implementations at various levels of detail, ranging from detailed biophysical models of ion channels, to much simpler integrate-and-fire models, to even simpler leaky integrator models. The appropriate level of granularity should be determined by the experimental data that the model is intended to address: The present work aims to (ultimately) address the behavior of embodied turn taking in terms of its directly underlying mechanisms and thus neural models at the level of ion channels may be inappropriately myopic. On the other hand, various linguistic tasks have been used in EEG and brain imaging experiments that provide a more global view of brain function. For the current task it seems that the most appropriate level of neural modeling must have modules

[2] Andrew Gargett, personal communication.

corresponding to relatively large-scale brain regions and include an account of action generation.

The neural model explored here, Augmented Competitive Queuing (ACQ), is a model related to basal ganglia and cortical interactions that allow context-dependent chains of actions to be learned and flexibly deployed. The model, developed by Bonaiuto and Arbib (unpublished), implements action selection among competing motor schemas in a parallel neural network. Each action's relative competitive weight depends on its desirability, which is learned via reinforcement learning (Sutton and Barto 1998). The desirability is the estimated value of action, which is the expected sum of future rewards, given that a particular action is executed in a particular context.

In this chapter we describe a hybrid model that integrates features of the YTTM and ACQ by expanding key cognitive components of the former with neural mechanisms from the latter. The model is able to learn turn taking with little or no overlap in speech and to learn "social" turn taking cues. Furthermore, a key quality of the model is its highly extensible framework. In four experiments we investigate how turn–taking behaviors emerge in the system and how different patterns of conversation unfold with various parameter settings. The results of simulation experiments on these neurally-implemented modules are reported and ongoing work on integrating these submodules into a more complete neural model of turn taking is described.

The chapter is organized as follows: First we give an overview of related work and describe briefly YTTM and ACQ. Then we detail the integration of the two systems, how a subset of the YTTM has been implemented as an ACQ module. We then explain the setup and results of four experiments where two identically structured, simulated agents interact: The first exposes general properties of the new turn–taking model; the second explores the convergence of turn–taking cues when more than two agents are trained in a round-robin manner, producing agents with a common ("socially shared") set of turn-taking behaviors; the third experiment tests the system at varying levels of motivation to speak, producing different patterns of turn negotiation. The final experiment replaces the simpler motivation signal with a more realistic form that yields interesting patterns of turn-taking behavior. Finally, the results of the experiments are discussed and further research directions are considered.

19.2 **Related work**

One of the most influential models of turn taking over the past 30 years has been that of Sacks *et al.* (1974)—a model focused on smoothness in interaction. While a significant achievement in the study of dialog, the model's focus on language and syntax has been criticized (Thórisson 2002; O'Connell *et al.* 1990), as has its lack of accounting for semantics and pragmatics as potential contributing factors to conversational (turn-taking) organization (O'Connell *et al.* 1990). While O'Connell *et al.*'s criticism of Sacks *et al.*'s and related work does not state so explicitly, their arguments point to the fact that multiple goals and complex constraint satisfaction (such as non-overlapping speech and high-level "Gricean Maxims" in general (Dale and Reiter 1996)) is often part of the goals of people engaged

in dialog: "The ultimate criterion for the success of a conversation is not the 'smooth interchange of speaking turns' or any other prescriptive ideal, but the fulfillment of the purposes entertained by the two or more interlocutors." (O'Connell 1990, p. 346). To this it can be added that as long as the clear and concise interchange of information is a goal of the participants—which it is in a significant portion of both casual and formal conversations—the avoidance of simultaneous speech, which obviously can lead to mis-hearings and misunderstandings, will also be one of their (sub-)goals. Conversely, where possible, the avoidance of long silences (which are sometimes perceived as "awkward") will speed up the exchange of information. Such goals, which clearly can dynamically change between (and even during) dialogs, will need to have a place in any model that wants to explain in general terms how the observed behavior patterns in dialog come about.

More recently, Iizuka and Ikegami (2002, 2004) describe a system with two interacting agents playing a game of tag. While not specifically addressing turn taking in human dialog, the research shows that various patterns of emergent turn taking can ensue depending on how the control systems in the robots are constructed. Notable in their work is the modeling of prediction mechanisms in the agents—an important factor in any theory that wants to explain real-time turn taking. Prediction is also the focus of Wilson and Wilson's work (2005) and Schlangen (2006). The former propose a coupled oscillator model to explain the tight coupling of interlocutors observed in real-time dialog. The latter showed how various machine learning techniques could reach human-level performance in predicting turn-holding and turn-giving using various features extracted from pitch and syntax. Sato *et al.* (2002) likewise found clear benefits of prediction. They used a learning algorithm to generate a decision tree that could predict and identify turn-taking points in simple Japanese office dialog. Using a data set containing detailed prosody analysis, word, word category analysis, and internal recognition/ understanding state of the system, their method achieved 83.8% accuracy. Although the dialogs were simpler than the average natural conversation (the speech recognition had a vocabulary of 161 words), these results point to the importance of taking multiple features into account to achieve natural turn taking.

The YTTM of Thórisson (1996, 2002) is a model of turn taking that addresses manual gesture, gaze, body stance, speech semantics, intonation, and the integration of these in a coherent manner, as well as the planning and delivery of coordinated gesture, facial expression, gaze, and speech content relevant to interaction in real-time dialog. The YTTM has been implemented for two-party, task-oriented conversations (Bryson and Thórisson 2000; Thórisson 1996) and shown to generate natural turns and multimodal behavior in a highly dynamic fashion in interaction with two kinds of gesture (deictic and iconic), continuous speech, indication of attention (body, head, and gaze direction), in a relatively unencumbered and natural manner. While many features of turn taking are still missing from the model, it takes semantics into account and is multimodal. It builds on several cognitive hypotheses about turn taking, some of which are discussed in this chapter.

As mentioned above, we view dialog interaction is an emergent property of complex interactions among cognitive processes—a complexity that work to date clearly has only begun to address (cf. Duncan 1972; Duncan and Fiske 1977; Goodwin 1981; Thórisson

2002; Wilson and Wilson 2005; Kopp *et al.*, this volume). Unlike O'Connell *et al.* (1990), therefore, we do not believe that language syntax is the "wrong" place to start explaining turn taking[3] any more than we believe information exchange is the "right" way to view or model dialog: We see a need to take both into account, as both information content and surface phenomena (e.g. intonation; cf. Grosjean and Hirt 1996) have been seen to affect turn taking and related behaviors in real-time dialog. To create systems that are capable of high interaction complexity, and to explain the interaction patterns observed in real-time human dialog, we have to undertake a fairly complex modeling effort.

To our knowledge, the neural mechanisms of turn taking have not been explored specifically, but the cognitive mechanisms needed for turn taking include perceptual processes, memory processes, and motor planning and control, all of which have been studied extensively in the last 30 years. Recent efforts to build complete models of cognitive skills involving the integration of all of these include goal-directed imitation (Erlhagen *et al.* 2006), navigation (Guazzelli *et al.* 1998), and conflict monitoring (Botvinick *et al.* 2001). The work of Bonaiuto and Arbib on ACQ (unpublished) provides an account of interacting perceptual and motor neuroschemas in generating flexible sequences of goal-directed actions. Further details of ACQ, as well as the YTTM, are given in the following sections.

19.2.1 YTTM

The YTTM (Thórisson 2002) is based on the Ymir model of cognition (Thórisson 1996, 1999) that models cognition as a set of interacting processes (Box 19.1). All of these play a role in the YTTM, although some are more important than others for the turn taking proper: If we assume a single conversational topic, content-related mechanisms, for instance, do not need to be explicated for understanding or even implementing a basic turn-taking system; they can be abstracted through very simple operating principles, as was done in the present study (Box 19.2). The operating assumptions about content understanding and generation are that they are incremental processes that can plan utterance content opportunistically as well as ahead of time.

YTTM proposes that (1) turn-taking mechanisms are fairly isolated from content interpretation and generation systems and that (2) the systems interact to coordinate the global activity of the body during conversation, via a set of primitives. This set of primitives is a relatively small one (Table 19.2). The YTTM further proposes that (3) turn-taking and content systems interface with a limited-resource planning system that serves both (Action Scheduler).

The split proposed in point (1) above is composed of two main categories of processes, envelope and content interpretation and presentation. Envelope processes, and resulting behaviors, are explicitly intended for managing the turns and are not related to the conversational topic; content interpretation/ presentation processes, and resulting behaviors,

[3] However, we *do* feel that proposing syntax as a general or primary (or – heavens forbid – the only) approach to modeling dialog phenomena is a dead end.

Box 19.1 The set of cognitive components proposed by the Ymir Turn-Taking Model (YTTM)

P, set of perceptual feature processes
D, set of decision-making processes
C_u, content understanding mechanism
C_g, content generation mechanism
B, behavioral displays
P, G, plans with goals
$P = \{p_1 \ldots p_n\}$
$D = \{d_1 \ldots d_n\}$
$B = \{b_1 \ldots b_n\}$
$G = \{g_1 \ldots g_n\}$
$P = \{p_1 \ldots p_n\}$

manage the topic of the conversation, and thus require knowledge of that particular topic. One argument for such a split comes from the observation that the interaction skills can, to some extent, be independent of all possible discussion subjects (it could itself of course be a topic of discussion, but such a discussion could not proceed without following the very rules being thus discussed). We find it unlikely that a unique set of dialog skills would exist for every topic or field of expertise that one could be proficient in. This echoes arguments heard from proponents of the massive modularity hypothesis of cognition (cf. Fiddick *et al.* 2000). Another argument is the law of parsimony: Evolution seldom favors a baroque solution over a minimalist one, given the choice.

Envelope behaviors are controlled through a set of modules (Deciders) with time-sensitive rules that are hierarchically organized in each participant. The hierarchy indicates precedence or priority of control; envelope processes and behaviors are of a higher priority than content processes and behaviors. Simple modules monitor and inform the more complex cognitive processes that participate in dialog: Memory, planning and execution, topic knowledge, etc., and their states. It is the interaction between these processes, via the connecting envelope modules (and their rules), that generates the behavior patterns observed in human dialog. Behaviors such as quickly gazing away and back when taking the turn (Goodwin 1981), lifting eyebrows when being asked a question, etc. are examples. Using the different priorities for envelope and content behaviors the Action Scheduler manages conflicts between the various plans, plan snippets and decisions, and helps coordinate them 2 to 4 seconds into the future. Decision-making modules link perceptions to actions in a way not unlike behavior-based AI (cf. Brooks 1986), however, the modules in the YTTM allow more indirect connection between sensing and acting as well as hierarchical constructs, and thus go beyond, for example, the subsumption architecture (Brooks 1986).

Box 19.2 Primitives of the YTTM connecting content management systems with the turn-taking management systems (indentation indicates subtypes of the type above)

Topic-System-Received-Speech-Data
  Speech-Data-Available-For-Content-Analysis
Topic-System-Interpreting-Perceptual-Data
  Topic-System-Interpreting-Speech-Data
  Topic-System-Interpreting-Multimodal-Data
Topic-System-Successful-Interpretation
Topic-System-Act-Available
  Topic-System-Communicative-Act-Available
  Topic-System-Realworld-Act-Available
I'm-Executing-Content-Communicative-Act
  I'm-Executing-Content-Multimodal-Act
  I'm-Executing-Content-Speech-Act
I'm-Executing-Content-Realworld-Task

19.2.2 ACQ

The inspiration for Augmented Competitive Queuing (ACQ) comes from a study of forelimb movements in cats. Alstermark *et al.* (1981) experimentally lesioned the spinal cord in order to determine the role of propriospinal neurons in these movements. These experiments also happened to illustrate interesting aspects of the cat's motor planning and learning capabilities. In particular, the reorganization of the cat's reach and grasp motor program after the lesion suggested that the program was composed of a set of interacting and competing motor schemas, rather than being based on some sort of higher-level cognitive mechanisms. ACQ emphasizes how motor plans may emerge through patterns of competitive queuing (Houghton and Hartley 1995; Bullock and Rhodes 2003) based on the dynamic updating of values acquired through reinforcement learning (Sutton and Barto 1998). A key difference between ACQ and "classical" competitive queuing (CQ) is that the activation levels of motor program elements are dynamically computed in each "time step", rather than being completely specified before sequence execution. This allows action sequences to emerge dynamically with elements of the sequence flexibly deployed rather than being rigidly iterated through.

At the core of the ACQ model is a network for internal state representation, action recognition, and action selection (Figure 19.1). Actions are selected by the parallel planning and competitive choice layers given the outputs of the internal state representation and action recognition module. The output of an adaptive critic provides an error signal to modify the network weights on the basis of an external reward signal and an efference copy of the currently executed action.

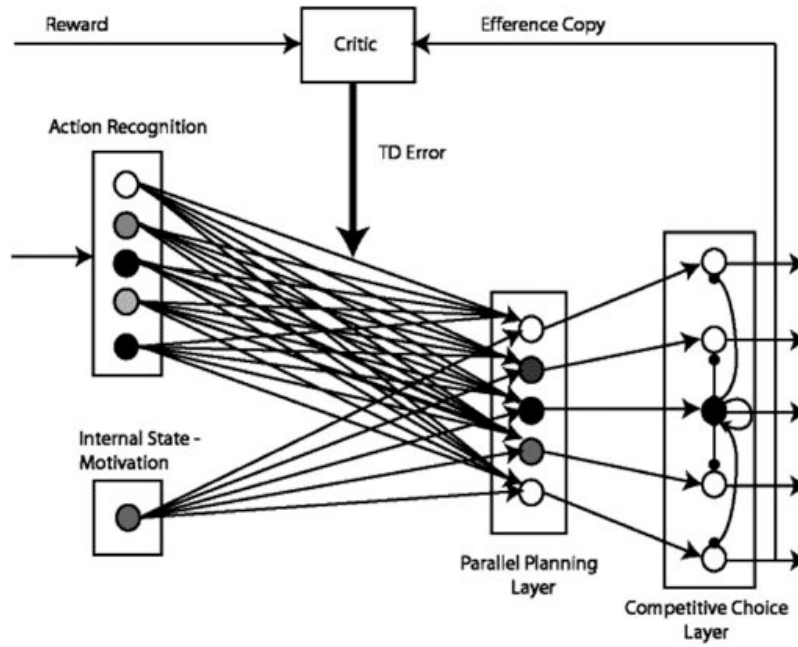

Fig. 19.1 The basic network for augmented competitive queuing (ACQ) at the core of the model's functionality. The activation of motor schemas in the parallel planning layer is composed of desirability values based on the output of the action recognition module and motivation signals from the internal state. For simplicity, only the lateral connections of one neuron in the competitive choice layer are shown. The other neurons in that layer have similar on-center, off-surround patterns of connectivity.

19.3 **YTTM and ACQ: architectural comparison**

The YTTM was developed using a precursor to the *constructionist AI* methodology (Thórisson *et al.* 2004), and ACQ was developed using *schema theory* (Arbib 1992). Both of these methodologies stress the decomposition of complex systems into hierarchically organized modules that interact through well-defined interfaces. Although YTTM and ACQ approach the problem of behavioral modeling from artificial intelligence and computational neuroscience respectively, the design methodologies used in each model make them amenable to comparison and integration. Both models use similar input–output mechanisms and the behaviors they produce are the result of complex interactions between relatively simple components. While certain core mechanisms of the two models may be interchangeable, there are significant differences, the main ones being *learning* (ACQ has it, YTTM doesn't) and *variable-duration actions* (YTTM has it, ACQ not). However, the modular decomposition facilitates a comparison of the models on a module-by-module basis, a clear advantage when combining relatively complex models like these.

The two models have similarities in perceptual input and action selection. YTTM distinguishes perceptual input models into unimodal and multimodal perceptors.

Unimodal perceptors receive input from a single mode only, for example hearing or vision, while the multimodal perceptors integrate information from unimodal and other multimodal perceptors. Context can drive the activity of perceptors—as they are functionally defined, their relevance to the current state of the agent can drive whether they are active or not and thus influence the information being extracted from the senses. ACQ uses the very compatible concept of perceptual schemas for input. These schemas may be unitary—signaling changes in a single perceptual feature, or may be further decomposable into a network of lower-level perceptual schemas. ACQ does not make an explicit distinction between unimodal and multimodal components: The former case directly corresponds to unimodal perceptors while the later case subsumes multimodal perceptors as a case of perceptual schemas with multimodal subschemas, thus adding an extra organizational component, namely the grouping of the perceptors into schemas.

The core ACQ network fulfills the basic functions intended by the Action Scheduler of YTTM (Thórisson 1997) while lacking the hierarchical element of action organization. Both models stress real-time mechanisms in action selection and planning. However, YTTM represents hierarchical, decomposable goals and subgoals and ACQ represents sequences of primitive actions directed toward an ultimate goal. A key feature in YTTM's scheduler component is that behaviors are dynamically scheduled at various levels of detail with the ability to arbitrarily trigger subgoals. Arbitrary triggering is also possible in ACQ but it does not allow scheduling at various levels of detail. Arbitrary triggering of subgoals requires an explicit representation of the current context that includes a nested representation of the currently selected goals and subgoals. In YTTM this is done by deciders—context-sensitive modules that monitor the agent's mental state and make decisions about overt or covert action (Thórisson 1998), while the Action Scheduler receives these goals and selects between morphologies for achieving overt actions that satisfy them (based on the state of the body at any point in time). Triggering of action in ACQ is based on the firing rate of artificial neurons but is in principle comparable to the YTTM decision mechanism. Both ACQ and YTTM can monitor the progress of subplans and do replanning, while only ACQ can learn alternatives to failed plans. Both approaches focus on short-term plans (2–4 seconds long).

The functional equivalence of the input modules of YTTM and ACQ and the correspondence between YTTM's scheduler and the core network of ACQ are sufficient for an initial integration of the two systems in a neural model of turn taking. However, the models differ in action duration variability, goal representation, and learning: YTTM uses time-stamping mechanisms to schedule actions of varying duration while in its current state ACQ represents the lowest-level actions as having a unit length duration.

19.4 System design: the hybrid model

The turn-taking system we have implemented focuses on the action scheduling and multimodal behavior perception. To accommodate the increase in detail that is required for an initial implementation using mechanisms from ACQ the present system strips

away much of the details of conversation and focuses on the emergence of turn-taking behavior in agents whose perception–action associations are learned and coupled together. In particular, we make the assumption that both agents have the goal to take turns efficiently by avoiding overlapping speech and silences. Since the speech in the present experiment does not contain any semantics, we further make the assumption that the Boolean speech signal only represents content-related speech, not envelope-related speech (back-channel feedback) or other speech functions.

The system comprises two agents, each consisting of an ACQ module for action selection (Figure 19.2). For the purposes of the present study the motivation, action recognition, and reward administration equations have been modified from the original model (see below). The remaining unchanged equations are also included below for completeness. Inspired by the J. Jr. system (Thórisson 1993), each agent is capable of three "speaking" actions, designated *speak-intonation-up*, *speak-intonation-flat*, and *speak-intonation-down*, as well as four extraneous actions: three oculomotor actions (*look-at-face*, *look-away*, *look-at-workspace*), and one manual action (*move-hands*). Since these actions have not been grounded in an embodiment yet (real or simulated) which would force a similarity with human use of these actions, the names for these non-speaking actions are not meaningful in the present experiments and thus are henceforth referred to as *speaking* and *non-speaking* actions, respectively.

Input to the ACQ module comes from two sources: Perception of the other agent's actions as well as an internal motivational signal that represents the desire to speak (*motivation-to-speak*). This motivation-to-speak signal replaces the original ACQ executability parameter, which in the original model gated the activity of action representations based on physical possibility. It is assumed that the output of ACQ projects to a lower-level motor control structure which is not modeled here. Likewise, the perceptual processes of the action recognition modeled are not modeled. Therefore, the recognition

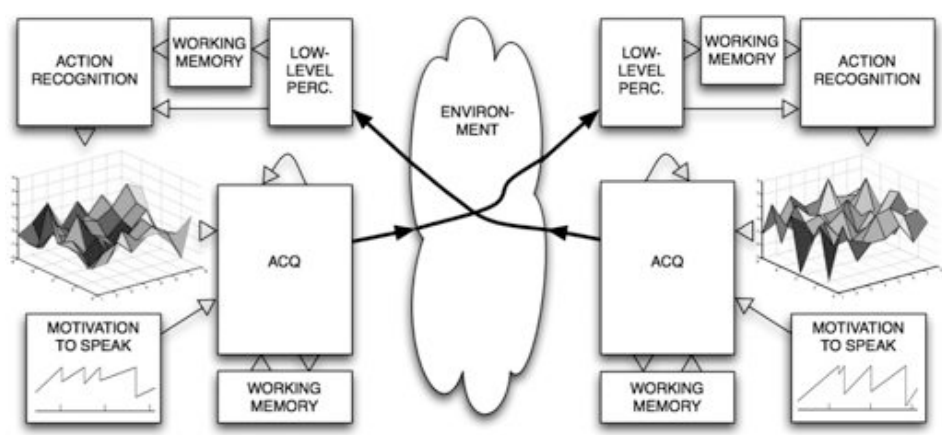

Fig. 19.2 The system setup consists of two identical systems capable of expressing multimodal behaviors, each with slightly different parameter settings for motivation to speak (see text for details).

of motor schemas is modeled by corrupting the output signal of one agent with noise and feeding it as perceptual input to the other agent.

The connection weights within the ACQ network between the action recognition module and the parallel planning layer are shaped via temporal difference (TD) learning (Sutton and Barto 1998), a form of reinforcement learning. The reinforcement signal is generated on the basis of the actions currently executed by each agent such that when one agent is speaking there is a positive reward signal, when neither agent is speaking there is a slightly negative reinforcement signal, and when both agents are speaking there is a strongly negative reinforcement signal. The idea is that the agents will learn to coordinate their internal motivational signals with the perception of the other agent's actions in order to maximize the reward.

Successful coordination of turns requires an element of prediction of the other agent's behavior (Schlangen 2006; Thórisson 2002). If both agents simply learn to speak when the other is not speaking, they could potentially oscillate between periods speaking simultaneously and silence (a behavior observed in prototype versions of the present model). If both agents start out silent, this naïve strategy would cause both agents to start speaking in the next time step. They would then simultaneously perceive the speech overlap and select a non-speaking action or no action for the next time step. Without the ability to predict the actions of the other agent, at least to some extent, the system is vulnerable to becoming trapped in such a cycle (note that this sometimes occurs when two humans begin speaking at the same time and simultaneously yields the turn). We investigate under what circumstances successful turn-taking behaviors emerge within the system and how different patterns of conversation unfold with various parameter settings.

19.4.1 Implementation details

19.4.1.1 System overview

At the core of this model is a network for motivational state representation, action recognition, and action selection (ACQ). The motivational state schema generates a signal that varies with time and indicates the urgency of speaking. The perceptual inputs are the recognized actions of the other agent. The ACQ module for action selection consists of two layers of processing units: a parallel planning layer and a competitive choice layer. Units in the parallel planning layer combine the perceptual inputs with the output from the motivational state schema. As in CQ, units in the parallel planning layer form direct excitatory synapses on corresponding units in the competitive choice layer. The competitive choice layer implements an intralayer WTA process. Each unit in this layer has an on-center, off-surround projection profile—it excites itself and inhibits surrounding units. The balance of excitation and inhibition ensures that the unit with the highest net excitation at each point in time will generally remain active while others will be inhibited.

19.4.1.2 Schema implementation

The behavior of the ACQ layers is described as projections between layers of leaky integrator neurons. Leaky integrators are artificial models of neurons that simulate their

mean firing rate based on axon hillock membrane potential. Membrane potential is assumed to be proportional to afferent input and a time constant derived from the membrane's capacitance and resistance. A saturation function is applied to the membrane potential in order to calculate the mean firing rate which is bounded by 0.0 and 1.0. It is common to use leaky integrators to model sets of interacting brain regions, rather than compartmental models which are commonly used for modeling small networks and single neurons.

While the leaky integrator neurons operate on a continuous time scale, the reinforcement learning algorithm used to modify the connection weights between them operates on an event-driven, discrete time scale. To distinguish between them, time in the continuous scale is labeled t and time steps in the discrete scale are labeled T. In a sequence of L actions, each having a continuous duration of D, the discrete time scale range is $1 \leq T \leq L$ and the continuous time scale range is $1 \leq t \leq (L-1)D$.

19.4.1.3 Motivation level module

In a complete system the motivational signal would consist of a combination of output from higher-level modules which would plan utterances, the emotional state, as well as the inferred internal state of the other agent. Our model greatly simplifies the motivational signal of agent i at time t, $mi(t)$ into a combination of two variables: $ai(t)$, agitation, and $hi(t)$, have-something-to-say. Rather than modeling higher-level cognitive modules for the perception, interpretation, and production of language, we use two models for agitation and motivation to speak. First, in Experiments I, II, and III, $hi(t)$ for agent i at time t is approximated as a sine wave with a given frequency and phase, bounded by 0 and 1:

$$h_i(t) = \frac{1 + \sin(\omega_i t + \phi_i)}{2}$$

where ω_i is the frequency, t is the time, and ϕ_i is the phase shift. The output of the motivational state module for agent i at time t is given by:

$$m_i(t) = \Theta\left[a_i(t) + \frac{h_i(t)}{2}\right]$$

The $\Theta(x)$ function is the saturation function:

$$\Theta(x) = \begin{cases} 0.0 & : \quad x < 0.0 \\ x & : \quad 0.0 \leq x \leq 1.0 \\ 1.0 & : \quad x > 1.0 \end{cases}$$

Thus at high levels of agitation, the motivation signal will saturate at 1.0. In accordance with YTTM, we separate speech and gesture content creation from its presentation via a set of primitives (Table 19.2). The present approximation is thus a placeholder, in order not to complicate the model and risk intractable results on first implementation. Second, in experiment IV, we created a more believable motivational model in which motivation

to speak dropped significantly after an agent had "said what it wanted to say", producing a sawtooth wave. We also removed the agitation component. The details of this function are provided below in Section 19.5.4 Experiment IV: "natural" motivation to speak.

19.4.1.4 Action recognition module

The action recognition neuroschema consists of an array of leaky integrator neurons which signal the recognition of the execution of an action by another agent. These neurons are the sensory input to the agent. As the low-level processes of perception are beyond the scope of this project, thus the input to the action recognition neurons is simply a copy of the action execution output of the other agent, corrupted by noise. Given N possible actions to execute, M of which are speaking actions (where M <N), $X_{i,y}(T)$ is equal to 1.0 if agent i executes action y at time T, and 0.0 otherwise (see Section 19.4.1.5 Action selection module, below).

The dynamics of the membrane potential of the action recognition neuron in agent i representing action y at time t, $u_{\hat{x}_{i,y}}(t)$ are given by

$$\tau_{\hat{x}_i} \frac{du_{\hat{x}_{i,y}}(t)}{dt} = -u_{\hat{x}_{i,y}}(t) + X_{i,y} + randn(\sigma^2_{\hat{x}_i}, 0.0)$$

Here, $\tau_{\hat{x}_i}$ is the time constant of the action recognition neurons in agent i, $\sigma^2_{\hat{x}_i}$ is the variance of the noise in action recognition in agent i, and $randn(\sigma^2, \mu)$ returns a normally distributed random number with mean μ and variance σ^2, and $1 \leq y \leq N$. The firing rate of the action recognition neuron in agent i representing action y at time t, $\hat{X}_{i,y}(t)$ is given by $\hat{X}_{i,y}(t) = \Theta[u_{\hat{X}_{i,y}}(t)]$, which bounds the firing rate by 0.0 and 1.0. We trained each agent with no noise in action recognition and in a series of simulation experiments determined the relationship between noise variance and action recognition error rate (Figure 19.6), and the effect of action recognition noise on turn taking (Figure 19.5).

19.4.1.5 Action selection module

The parallel planning layer integrates input from the motivational signal and the perceptual signal representing the recognized actions of the other agent. It is implemented as an array of N leaky integrator neurons, one for each motor schema (where N is the number of available actions). The firing rate of each neuron encodes the priority of the motor schema it represents. The dynamics of the membrane potential of the parallel planning layer neuron in agent i representing the action y at time t, $u_{PP_{i,y}}(t)$, are given by

$$\tau_{pp_i} \frac{du_{pp_{i,y}}(t)}{dt} = \begin{cases} -u_{pp_{i,y}}(t) + m_i(t)\left[\sum_{z=1}^{N} W^{\hat{X}}_{i,y,z}\hat{X}_{i,z}(t)\right] + randn\left(\sigma^2_{pp}, 0.0\right) & : 1 \leq y \leq M \\ -u_{pp_{i,y}}(t) + \left[\sum_{z=1}^{N} W^{\hat{X}}_{i,y,z}\hat{X}_{i,z}(t)\right] + randn\left(\sigma^2_{pp}, 0.0\right) & : M < y \leq N \end{cases}$$

where τ_{pp_i} is the time constant of the parallel planning layer neurons in agent i, $W^{\hat{X}}_{i,y,z}$ is the connection weight from the action recognition neuron representing action y to the

parallel planning layer neuron representing action z in agent i, and σ^2_{pp} is the variance of the noise in the parallel planning layer in agent i. Thus the motivation signal, $mi(t)$, only modulates the speaking actions, $1 \leq y \leq M$. The multiplicative combination of internal state and desirability of motor schemas by the parallel planning layer restricts motor schema competition to only non-speaking actions when internal motivation is 0.0 and biases action selection toward speaking or non-speaking depending on other values.

The random component of activation in the parallel layer ensures that a random motor schema is selected if neurons have similar levels of excitation. This introduces a level of exploration into each agent's behavior, without which new combinations of actions could not emerge. The noise is independent for each mode (speech, gesture, and gaze).

The firing rate of the parallel planning layer neuron in agent i representing the action y at time t is given by $PP_{i,y}(t) = \Theta[u_{pp_{i,y}}(t)]$. Each neuron in the parallel planning layer projects to a corresponding neuron in the competitive choice layer. Neurons in the competitive choice layer additionally receive lateral inhibition from the other competitive choice layer neurons, and self-excitatory input. The dynamics of the membrane potential of the competitive choice layer neuron in agent i representing action y at time t, $u_{cc_{i,y}}(t)$ are given by

$$\tau_{CC_i} \frac{du_{CC_{i,y}}(t)}{dt} = -u_{CC_{i,y}}(t) + pp_{i,y}(t) + \left[\sum_{z=1}^{N} W^{CC}_{i,y,z} CC_{i,z}(t-1) \right]$$

where τ_{CC_i} is the time constant of the competitive choice layer neurons, $W^{CC}_{i,y,z}$ is the connection weight from the competitive choice layer neuron representing action y to the one representing action z, and $CC_{i,z}(t)$ is the firing rate of the competitive choice layer neuron representing action z at time t, which is given via the saturation function, $CC_{i,z}(t) = \Theta[u_{CC_{i,z}}(t)]$. The connection weights, $W^{CC}_{i,y,z}$, form an on-center, off-surround connection profile which implements a winner-take-all process:

$$W^{CC}_{i,y,z} = \begin{cases} 1.1 & : y = z \\ -0.7 & : y \neq z \end{cases}$$

An action is selected for execution if the firing rate of the competitive choice layer neuron representing it is greater than a threshold, ε_i, and greater than twice the firing rate of every other competitive choice layer neuron:

$$X_{i,y}(T) = \begin{cases} 1.0 : \left[CC_{i,y}((T-1)D) \geq \varepsilon_i \right] \wedge \\ \qquad \left[CC_{i,y}((T-1)D) \geq 2CC_{i,z}((T-1)D), \forall z : y \neq z \right] \\ 0.0 : \textit{otherwise} \end{cases}$$

19.4.1.6 Learning

Each motor schema's desirability given the recognized action of the other agent is represented by the weights of the connections between the action recognition neurons and

those of the parallel planning layer and modified through temporal difference (TD) reinforcement learning. Temporal difference learning is done on the last executed motor schema based on the difference between its desirability and that of the currently executed motor schema. The motor schema currently being executed is determined by an efferent copy of motor signal, which is maintained as working memory trace. Learning takes place on the discrete time scale $1 \leq T \leq L$.

If both agents are speaking, both agents receive a reinforcement signal, $r(T)$, equal to –1.0. If one agent is speaking and the other is not, successful communication is rewarded by administering a reward signal of 1.0 to both agents. Silence is punished by administering a reinforcement signal of –0.1 to both agents when neither is executing a speaking action. The magnitude of each reward is arbitrary, however the relative values were chosen on the basis of prototype simulations.

$$r_i(T) = \begin{cases} 1.0 & : \left[X_{i,y}(T) = 1.0 \wedge \hat{X}_{i,z}(T) \geq \varepsilon_i\right] \wedge \left[(1 \leq y \leq M)\, XOR\, (1 \leq z \leq M)\right] \\ -0.1 & : \left[X_{i,y}(T) = 1.0 \wedge \hat{X}_{i,z}(T) \geq \varepsilon_i\right] \wedge \left[(M < y \leq N) \wedge (M < z \leq N)\right] \\ -1.0 & : \left[X_{i,y}(T) = 1.0 \wedge \hat{X}_{i,z}(T) \geq \varepsilon_i\right] \wedge \left[(1 \leq y \leq M) \wedge (1 \leq z \leq M)\right] \end{cases}$$

The following formulation is based on that of Sutton and Barto (1998).

$$\delta i(T) = \begin{cases} r_i(T-1)\left[\frac{1}{2} - m_i(T-1)\right] + y_i W^{\hat{X}}_{i,y1,z1} - W^{\hat{X}}_{i,y2,z2} & : \begin{array}{c}\left[r_i(T-1) = 1.0\right] \wedge \\ \left[X_{i,y}(T-1) = 1.0\right] \wedge \\ \left[M < y \leq N\right]\end{array} \\ r_i(T-1) + y_i W^{\hat{X}}_{i,y1,z1} - W^{\hat{X}}_{i,y2,z2} & : \quad otherwise \end{cases}$$

Here γi is the discount rate for future rewards for agent i, $z1$ is the action the agent i executed in this time step ($1 \leq z1 \leq N$, $X_{i.z1}(T) = 1.0$), $z2$ is the action the agent i executed in the previous time step ($1 \leq z2 \leq N$, $X_{i\ z2}(T-1=1.0)$, $y1$ is the action recognition neuron most active in the current time step ($1 \leq y1 \leq N$, $\hat{X}_{i,y1}(TD) \geq \varepsilon_i$), and $y2$ is the action recognition most active in the previous time step ($1 \leq y2 \leq N$, $\hat{X}_{i,y2}((T-1)D) \geq \varepsilon_i$). The motivation factor is included when the agent is silent and the other agent is perceived to be speaking to ensure that the effective reward for passive listening is inversely proportional to an agent's motivation to speak. This value is then used to update the desirability of the action executed in the previous time step:

$$\Delta W^{\hat{X}}_{i,y,z2} = \alpha_i \delta_i(T)$$

where α_i is the learning rate of agent i.

19.5 Experiments

Four experiments were performed using the Hybrid Model. Experiment I focused on general behavior of the system. We performed general tests to see whether the

system: (a) could learn turn taking; (b) responded correctly to systematic variations of parameters; and (c) provided enough flexibility to serve as a platform for further experimentation.

Having verified that the system operated according to expectations we performed a follow-up experiment, Experiment II, intended to see whether two dialog participants would develop the equivalent of common methods of displaying "turn signals", that is a common set of actions that help them take turns without speech overlaps, that is *content delivery* overlaps. (As the YTTM proposes a separation of envelope control from content interpretation and generation, the turn-taking behavior of the system defined by the observed speech patterns should be interpreted as representing the delivery of content-specific information only, not verbal delivery serving other functions, e.g. envelope feedback.) This was done by first running three simulated agents in round-robin interactions with each other. The hypothesis we wanted to test was whether, by interacting repeatedly with each other over a period of time, the system would settle on a common, shared set of turn-taking cues.

Subsequently we wanted to analyze further the exact nature of the turns produced by the system, so we selected two of the agents from Experiment II and ran two additional experiments: Experiment III examined the activity of the system when the agents had various levels of agitation. This increased the range of their motivational signal (motivation to speak), in turn increasing the probability that speaking actions would be selected. By increasing the agitation of both agents, we forced the already-trained agents to "confront" each other.

In the final Experiment IV we wanted to see the effects in our model of a more realistic motivation to speak. We hypothesized that in natural dialog the motivation to speak is sawtooth-shaped: In the canonical case a listener's motivation to speak will rise slowly as the speaker[4] keeps speaking, until either she is done speaking or impatience gets the better of the listener and he interrupts; at that point his motivation reaches a plateau that is held until he is done delivering what he wanted to say, at which point the motivation drops instantly.

19.5.1 Experiment I: baseline

Experiment I consisted of a series of pilot test intended to verify that all subsystems in the setup performed to specifications, and that the system could learn to interact. The setup for it was as follows: Two agents were trained while interacting with each other over 1000 trials, where each agent's motivational signal was represented with a sine wave. The agents had slightly different frequencies and phase offsets for their motivational signals.

19.5.1.1 Results

The results verified that the perception, action selection, and learning mechanisms worked correctly. The agents learn to take turns, each using a unique set of action recognition–action selection pairings. (See Figures 19.3 to 19.7 for details.)

[4] As already mentioned, we use the terms "speaker" and "listener" for convenience – more accurate terms for these roles are "content presenter" and "content interpreter" (see Thórisson 2002).

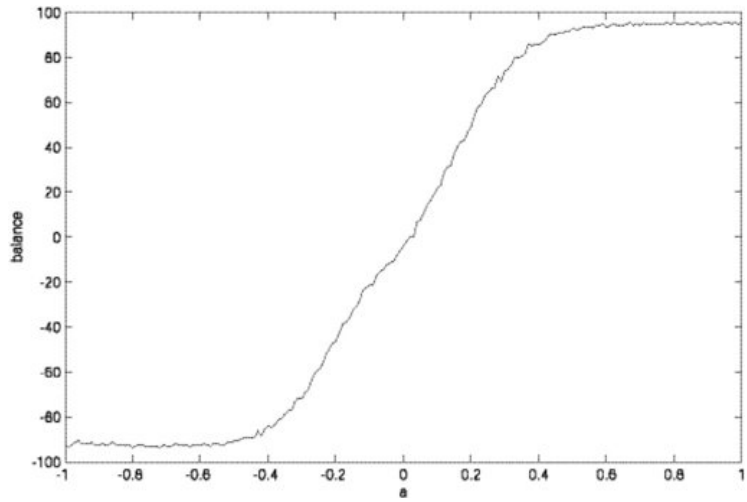

Figure 19.3 Conversational balance as a function of agitation. To test the effects of agitation on conversational balance, the agitation (a, x-axis) of agent 2 was set to 0.0, while the agitation of agent 1 was varied from 0.0 to 1.0. Then the agitation of agent 1 was set to 0.0, while the agitation of agent 2 was varied from 0.0 to 1.0. During each condition, the balance of the conversation was measured. In this figure the mean conversational balance during a conversation between agents 1 and 2 with various agitation settings is shown. The magnitude of negative agitation settings correspond to that of agent 2 when agent 1's agitation is held at 0.0, while positive agitation settings correspond to that of agent 1 while agent 2's agitation is held at 0.0. The results show a fairly predictable response that the system exhibits when maximum level of motivation-to-speak and up from 0.5 (that it was set at during training) to 0.75, for either agent. The results are a sanity check but also provide evidence that the system does not display any non-linearity related to motivation-to-speak.

19.5.2 Experiment II: social turn-taking signals

In our first experiment we found that when training two of our agents to interact they would learn to avoid content delivery (speech) overlaps. However, because each has an independent perception and action module, each would tend to learn its own unique set of actions to signal its state. In other words, the system did not show any sign of a common set of behaviors that could be compared to the "turn signals" observed in human dialog. Hypothetically, such a set of actions could be developed in our system—the equivalent of "social turn-signals": standard methods of behaving shared in a team that would help the agents converse with anyone on the team without speech overlaps.

We hypothesized that to get this effect we would have to train a minimum of three agents together, which should result in the emergence of a common set of actions (and ways to perceive them) that works across the "social population". The number of speaking and non-speaking actions was high enough in relation to the number of agents that there was the possibility for unique action associations between individual pairs of agents. On the other hand, given enough exploration the population could possibly converge on a common set of turn-taking cues (embodied in action recognition–production associations).

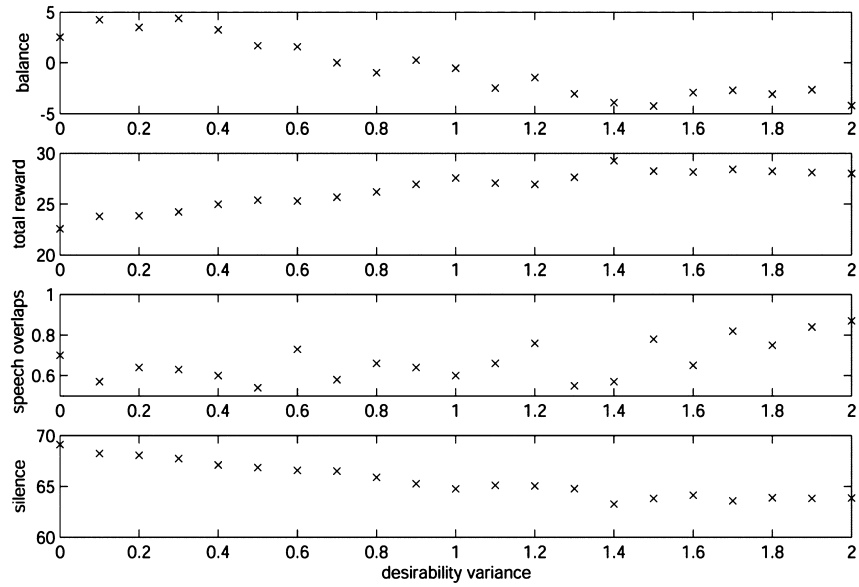

Figure 19.4 Balance, total reward, speech overlaps, and silence as a function of variance in desirability of action execution. Noise in the parallel planning layer was varied from 0.0 (no noise) to 2.0 in order to explore its effect on conversation. For each variance setting, the mean conversational balance, total reward, silence, and speech overlaps were averaged over 100 trials of 100 discrete time steps each. The variance of the noise in the parallel planning layer was varied from 0.0 (no noise) to 2.0 in order to explore its effect on conversation. In this figure, for each variance setting, the mean conversational balance, total reward, silence, and speech overlaps were averaged over 100 trials of 100 discrete time steps each.

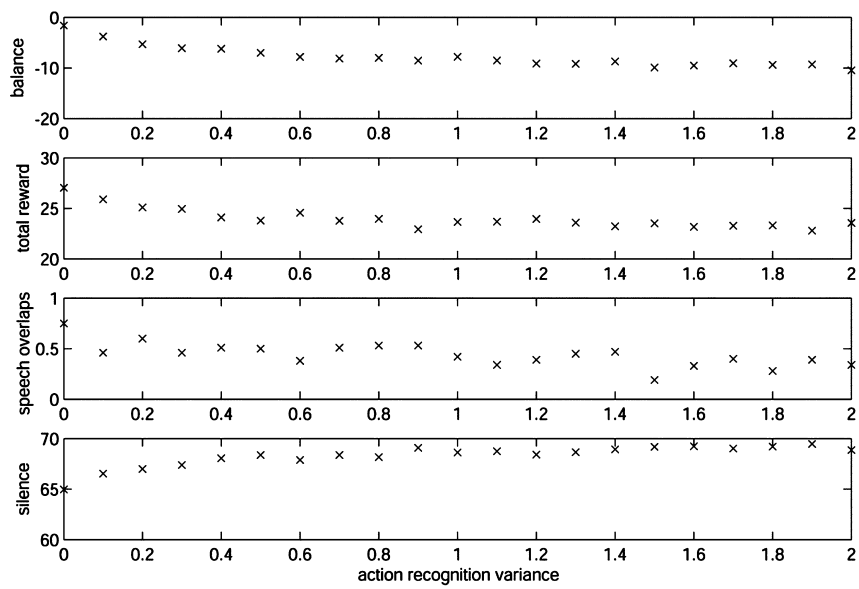

Figure 19.5 Balance, total reward, speech overlaps, and silence as a function of noise variance in action recognition. The variance of the noise in the action recognition module output was varied from 0.0 (no noise) to 2.0 in order to explore its effect on conversation. For each variance setting, the mean conversational balance, total reward, silence, and speech overlaps were averaged over 100 trials of 100 discrete time steps each. (From top to bottom: mean conversational balance, total reward, speech overlaps, and periods of silence averaged over 100 conversations between two agents.)

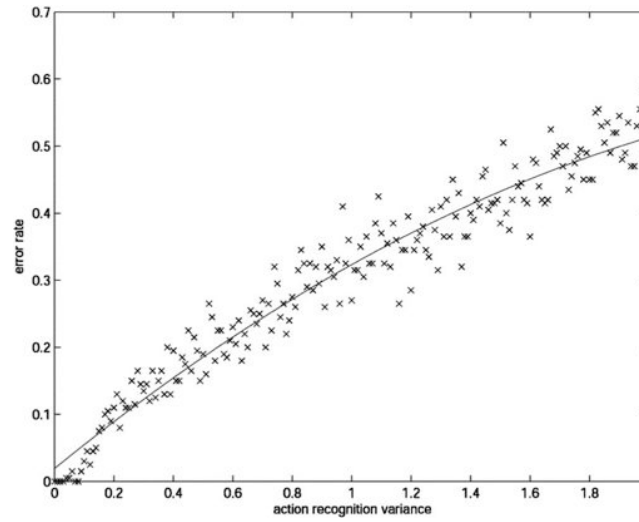

Figure 19.6 Action recognition error rate as a function of noise in action recognition membrane potential. The rate of action recognition error as a function of the variance of the noise in the action recognition membrane potential, $\sigma^2_{\hat{X}}$. The crosses denote the data points and the line depicts a 2nd order polynomial curve fitted to the data. Note that the polynomial only fits the data for $\sigma^2_{\hat{X}} > 0.2$. For $\sigma^2_{\hat{X}} \leq 0.2$ the actual error rate is lower than the theoretically predicted values.

The three simulated agents were trained in a round-robin fashion over 10,000 trials. At the start of each trial, two random agents were chosen from the set of three to run the simulation on. The weights of each agent were persisted from trial to trial.

19.5.2.1 Results

The results showed that the matrix representing the connection weights between the perception module and the action selection module converged in all three agents to

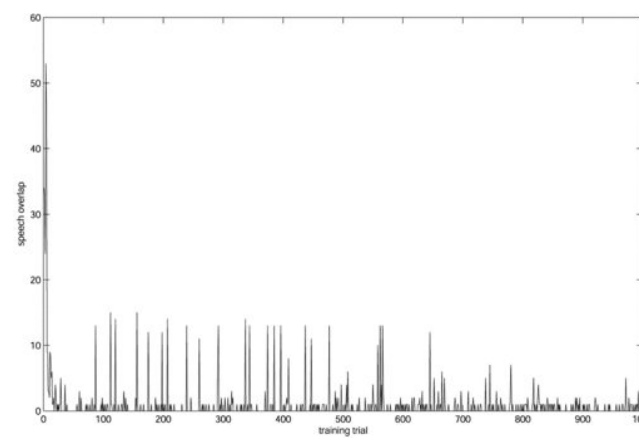

Figure 19.7 Number of speech overlaps per training trial (N = 1000) between two agents. The number of speech overlaps very high in the first trials (over 50) but rapidly decreased in subsequent trials (by less than 50). The next approximately 600 trials have a high variability in overlap number, until the variance decreases at around the 700th trial.

a very similar state, indicating that the agents were indeed learning a common set of action observation–execution associations. Figure 19.8 shows the mean Euclidian distance between the weight matrices of the connections between the action recognition and action selection modules of each agent when initialized with random values. The Euclidian distance between the two matrices of agents i and j was calculated by using the Frobenius norm of the difference between the two matrices ($\|W^{\hat{X}}_{i,y,z} - W^{\hat{X}}_{i,y,z}\|_F$). The Frobenius norm for a matrix A, $\|A\|_F$, is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}^2}$$

where $A \in \mathbb{R}^{nxn}$. Figure 19.9 shows how the matrices of all three agents appear before and after training.

19.5.3 Experiment III: high motivation to speak

To explore further the patterns produced by the system we changed the motivation of the "social" agents from Experiment II, increasing one while keeping the other constant. We also ran a simulation where the agitation of both agents was very high, so their motivation to speak was equal, and quite a bit higher than it had been on average in the training. In particular, we fixed agitation settings at $a = 0.5$, $a = 0.8$, and $a = 2$; have-something-to-say would vary (as a sine wave with frequency 0.1 for agent 1 and 0.008 for agent 2 and phase offset 0.0 for agent 1 and $\pi/2$ for agent 2) between 0 and 0.5 in all instances. Since the motivation-to-speak variable is equal to a saturation function applied to the sum of the agitation and have-something-to-say signals, it varies from 0.5 to 1.0 with

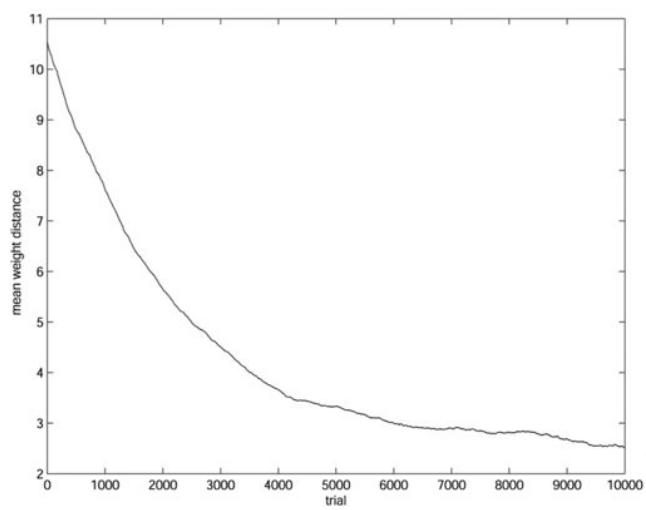

Figure 19.8 Mean Euclidean distance between the weight matrices of the connections between the action recognition and action selection modules in each of the three agents. The falling distance shows that each agent is learning to match the other's expectations about which non-speaking actions signal the agent's motivation-to-speak status.

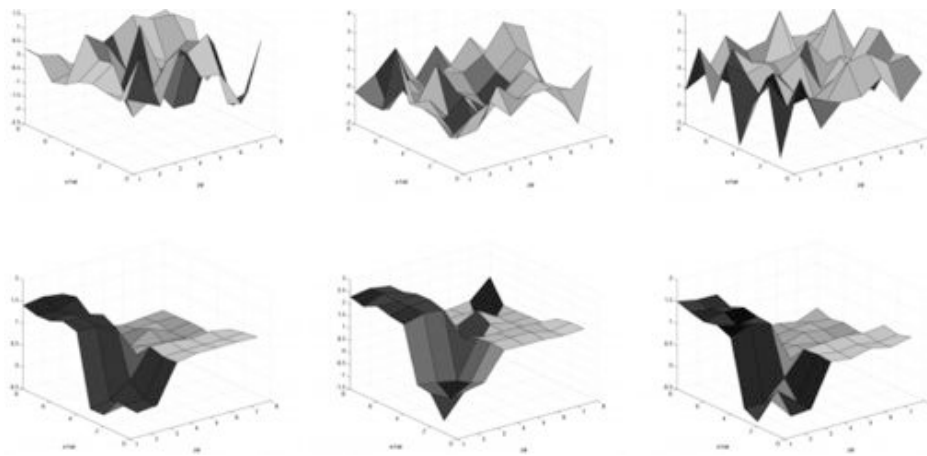

Figure 19.9 Matrices connecting action recognition module and action selection module in each of the three agents, plotted for each agent in three dimensions, before (upper row) and after (lower row) training, where convergence can be clearly seen in the highly similar final shapes of the matrices. (Upper row: Matrices of agents 1, 2 and 3 before training; lower row: Matrices of agents 1, 2 and 3 after training. X-Z-axis: Cell grid; Y-axis: Cell weight. Shading is random.)

$a = 0.5$, 0.8 to 1.0 with $a = 0.8$, and remains constant at 1.0 with $a = 2.0$. We ran 20 simulations of 100 ticks each with this setup.

19.5.3.1 Results

By and large the results we got were not very varied: Most of the time one agent would grab the floor and speak the whole time. For comparison, two baseline runs are shown in Figure 19.10A and B. In (A) the agents start speaking at the same time, then alternate bursts of speaking with one more time step of speech overlap. In (B) there are no speech overlaps and both agents speak at various points in the conversation.

Figure 19.11 shows another pattern observed. When looking at these graphs it may seem like the agents are negotiating turns through a sort of rock-scissors-paper game, with certain non-speaking actions acting as yielding signals. However, the mechanism that manifests itself this way is that for a particular agent the desirability of a non-speaking action is slightly positive when the other agent is speaking but random if the other agent is *not* speaking; desirability of speaking is positive when the other agent is not speaking and negative given that the other agent is speaking.

19.5.4 Experiment IV: 'natural' motivation to speak

Based on Experiment III we felt that the artificial nature of the sine wave motivation-to-speak might be producing unnatural speech patterns. We hypothesized that in natural dialog motivation to speak is in some cases sawtooth-shaped: Assuming an incremental construction of a response during listening, a listener's motivation to speak will rise slowly as the speaker keeps speaking, until either she is done speaking or the listener interrupts; at that point his motivation reaches a plateau that is steady and relatively

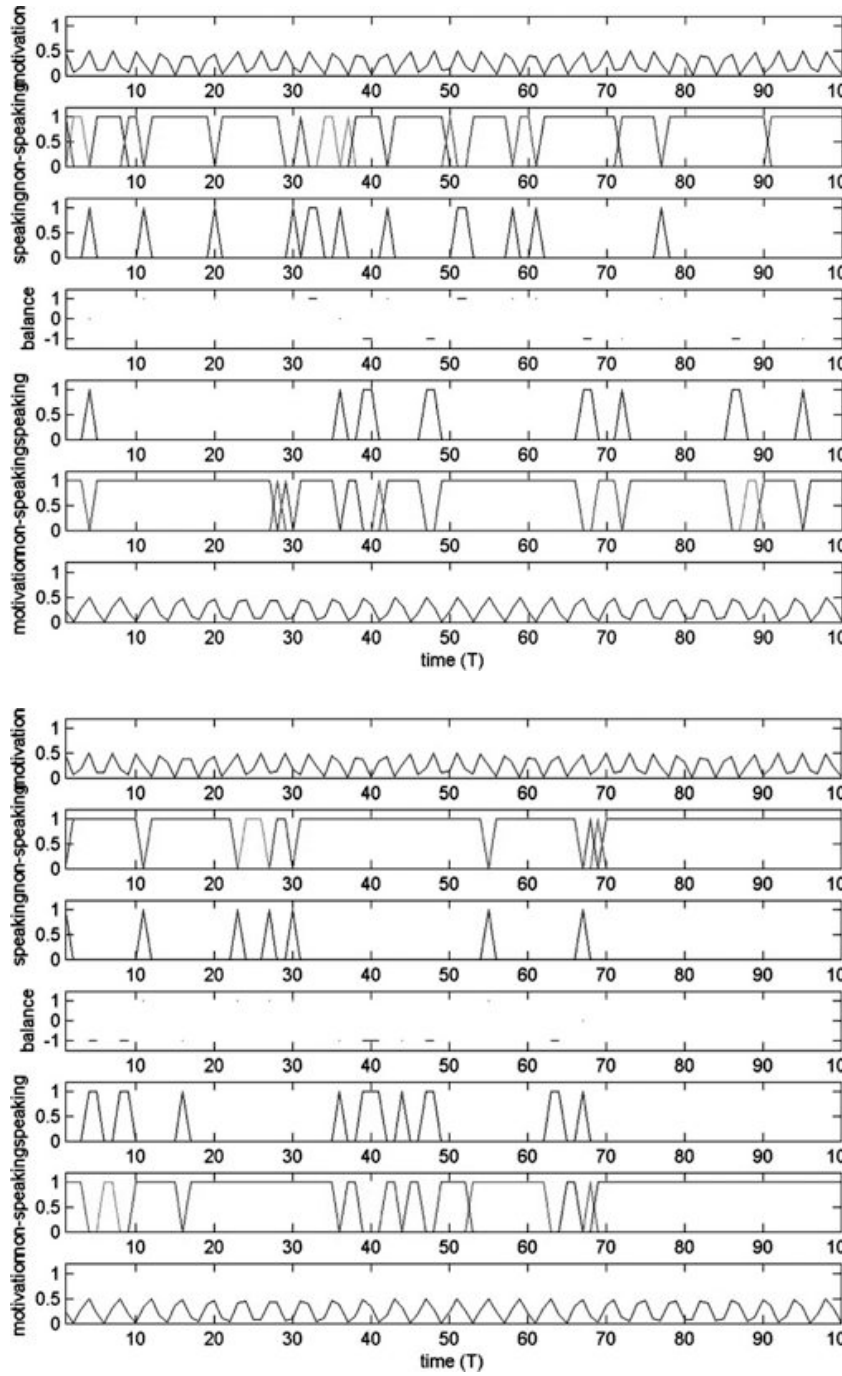

Figure 19.10 Baseline interaction for "socialized" agents. The following plots from two example runs show that at baseline, motivation-to-speak for either agent never rises above 0.5. This value is multiplied by the speak desirability when computing action priority, thus while both agents speak at different times, there are large periods of silence. (Agent 1 above middle box, agent 2 below: Center box shows who is speaking, with line in middle showing speech overlap and absence of line showing silence; boxes immediately above and below the center box show the speaking behavior of each agent, non-speaking behaviors are the second-to-top and second-to-bottom; top and bottom box plots motivation for each agent.)

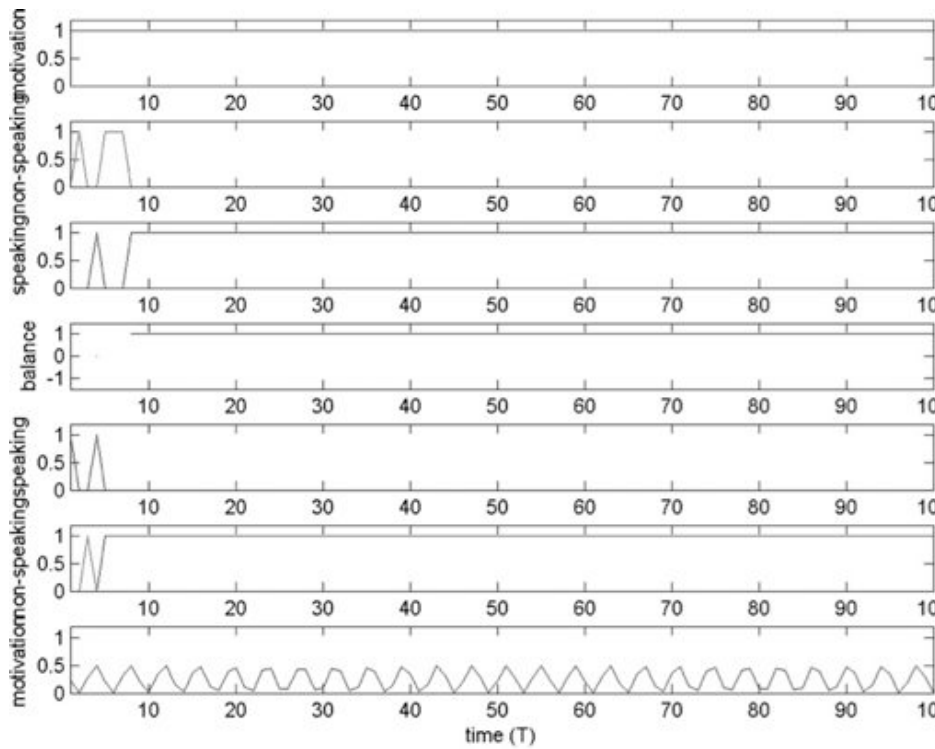

Figure 19.11 Negotiating the turn. In this trial agent 2 (bottom half) begins the session by speaking, in the first timestep. In the second time step agent 1 responds with a non-speaking action. After that both are silent for one step, then both start to speak at the same time. Both agents then switch to non-speaking actions on time step 5, but after that agent 1 starts speaking again and dominates the conversation for the rest of the period. As the initial simultaneous non-speaking action on both sides is initially the same, but on timestep 5 agent 2 switches to another kind of non-speaking action. Here agent 1 has an agitation level of 2.0 which causes its motivation to saturate at 1.0.) This pattern was found in about 15% of the initial part of sessions where motivation was very high in either or both agents. (Agent 1 above middle box, agent 2 below: Center box shows who is speaking, with line in middle showing speech overlap and absence of line showing silence; boxes immediately above and below the center box show the speaking behavior of each agent, non-speaking behaviors are the second-to-top and second-to-bottom; top and bottom box plots motivation for each agent.)

high until he is done delivering what he wanted to say, at which point the motivation drops instantly.[5] We created a motivation-to-speak signal that approximates this pattern and ran 20 new trials (of 100 steps each, same two agents) where the form of motivation-to-speak was a sawtooth wave with a different and variable frequency for each agent. The have-something-to-say signals were generated entirely independently based on different frequencies and phase offsets. At random points during the conversation the frequency was shifted by a random amount, but remaining at levels between 0.0001 and 0.0008. The low frequency was selected for stability and to allow the signal to reach its maximum value. The shift of the have-something-to-say signal frequency was added

so as not to get artifacts through arbitrary correlations between the initial states of each agent or phase correlations. We removed the agitation component of the motivation signal and removed the coefficient from the have-something-to-say signal: $m_i(t) = \Theta[h_i(t)]$. Given frequency ω and phase offset ϕ, the have-something-to-say signal at time t is given by:

$$h(t) = \omega\left[\left(\frac{t}{dt} + \frac{\phi\omega}{\pi}\right)\%\frac{1}{\omega}\right]$$

19.5.4.1 Results

The results show a clear turn-taking pattern and negotiations of the turn (Figure 19.12A to F). As can be seen, in the figures, a speaking party does not have to have maximum or even stable motivation to keep the turn during speaking. This is because there is some inertia to switch to another action once an action has been chosen. This is manifested in ACQ as hyperpolarization of inactive neurons in the competitive choice layer. Because of the saturation function on the firing rate, new inputs must raise the membrane potential of a neuron above 0.0 in order to affect its firing rate and make it eligible to potentially win the competition. If the current winning neuron sufficiently hyperpolarizes another through lateral inhibition, this imposes a lower limit on the intensity of input to the other neuron required in order for it to influence the state of the network. Speech overlaps are very rare in these runs, representing only about 0.45% of the total talk time, indicating that the system is negotiating turns very well. It seems that most, if not all of the non-speech actions serve some sort of "yield signal" (or "inverse interrupt"), as these tend to be on during periods where an agent is not speaking, and turn off just before the turn is switched.

When interpreting these plots, note that the agents only have a perception of the *last* recognized action of the other agent when planning their own *current* action. Thus, real predictive turn taking would ideally be seen when a speaking agent stops speaking after the other agent executes a particular non-speaking action for one step. Of course, however, the neurons in the network have a stochastic element to their membrane potential equations and the speed at which leaky integrators "charge" up is determined by their time constants, so the absence of this phenomenon does not indicate the absence of turn taking at all.

19.6 Discussion

In our model it seems that conflicting internal signals between two agents can be coordinated through a neurally-implemented perception–action learning mechanism. As can be seen in the data, the system will learn turn taking after the typical 5000–10000 training

[5] Clearly other patterns could be proposed, based on the mental state of the listener; an example would be the speaker asking a question and the listener not knowing how or what to say in response. In this case the motivation to speak would be fairly low. If, however, the listener really wanted to say something, but had not made up her mind about what to say specifically, she might use a standard pattern such as "Well, that's a good question". These kinds of patterns are best explored with a more accurate model of comprehension and content production.

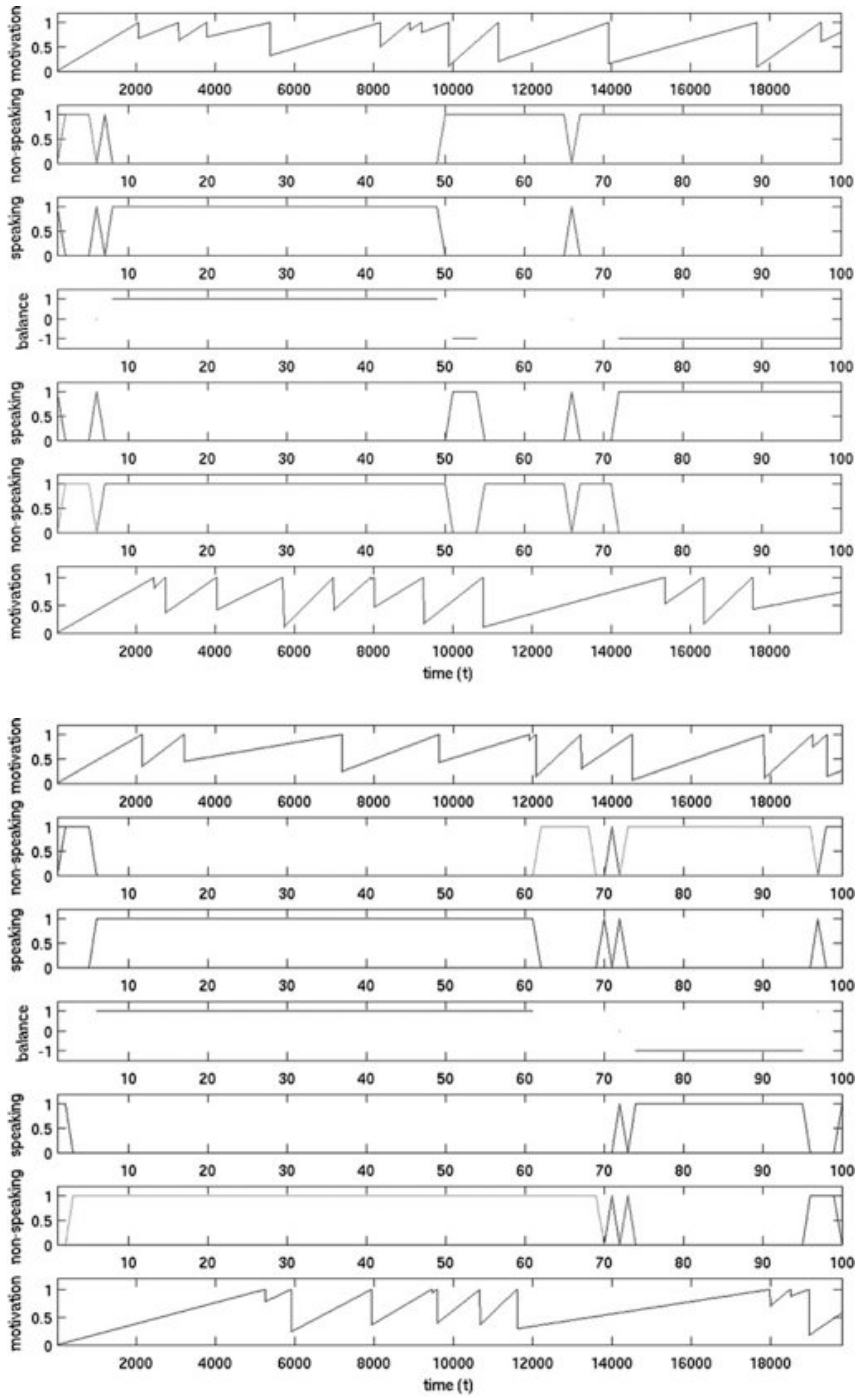

Figure 19.12A–F Example runs showing the use of turn signals in the agents. As the motivation-to-speak periodically drops significantly and grows linearly (random periodicity) in each agent, their only way to achieve smooth turn transitions is to develop common signals in non-speaking modes. (Agent 1 above middle box, agent 2 below: Center box shows who is speaking, with line in middle showing speech overlap and absence of line showing silence; boxes immediately above and below the center box show the speaking behavior of each agent, non-speaking behaviors are the second-to-top and second-to-bottom; top and bottom box plots motivation for each agent.

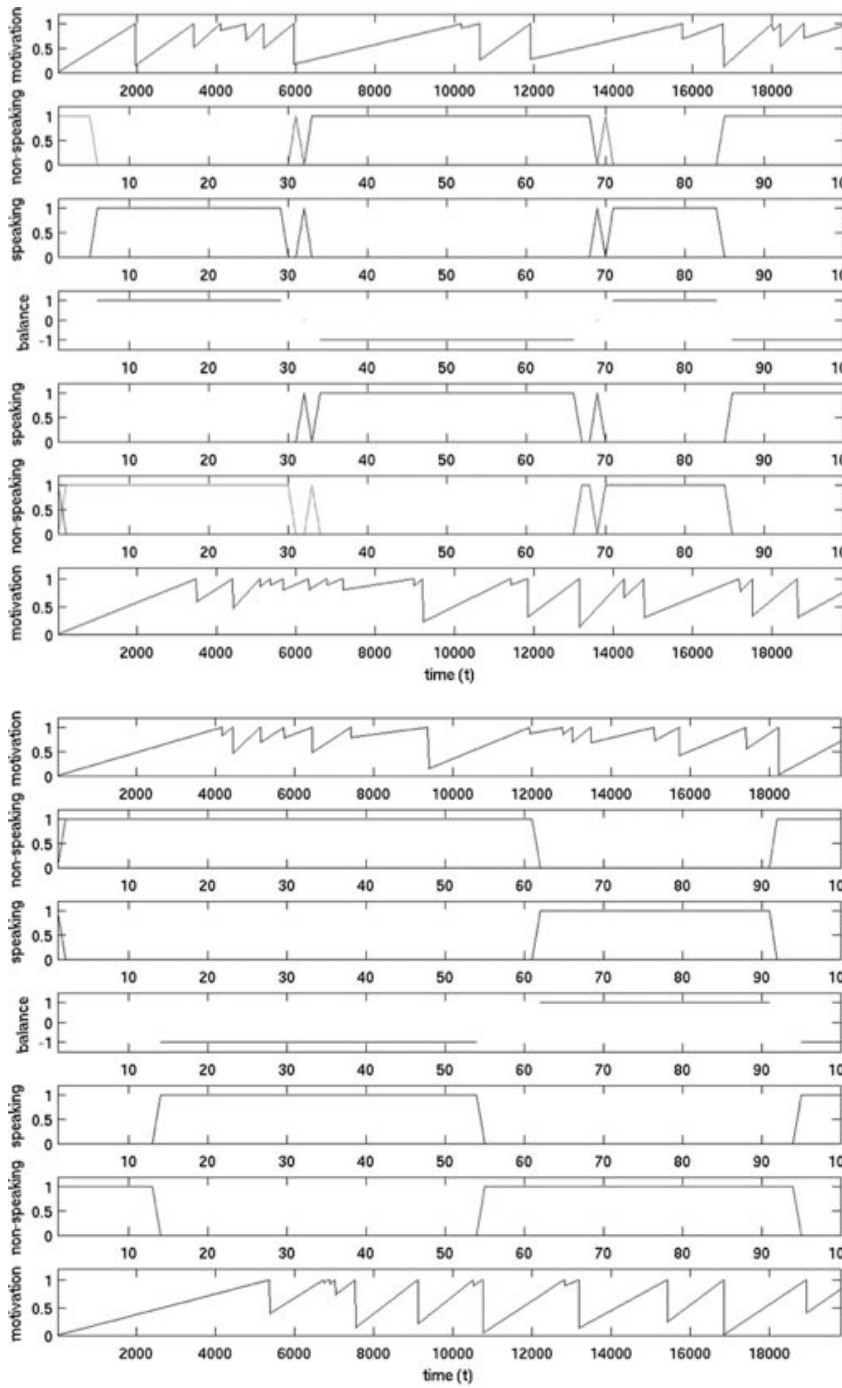

Figure 19.12A–F *(Continued)*

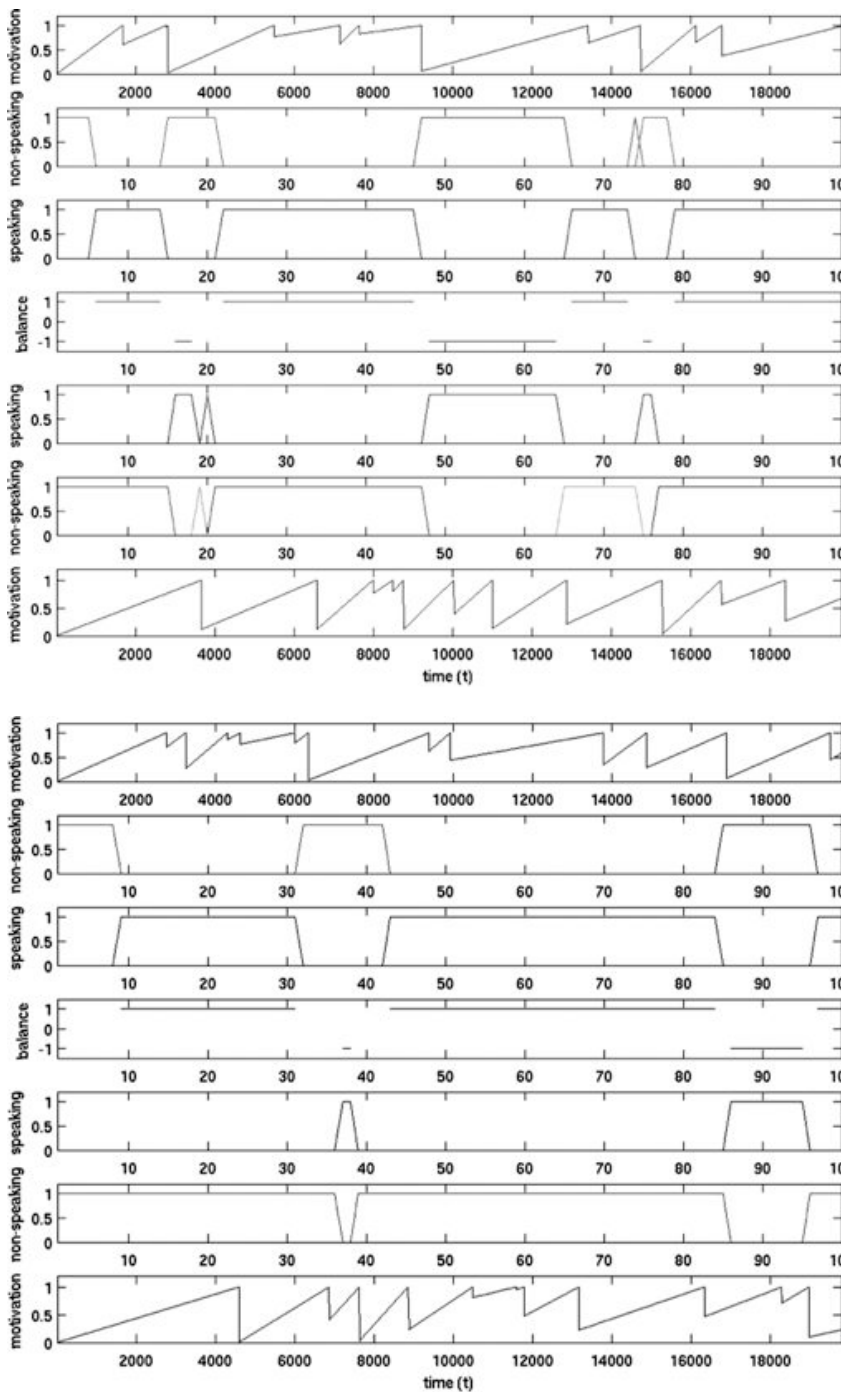

Figure 19.12A–F *(Continued)*

sessions that such systems require. When two agents are trained together, they each develop their own distinct set of action recognition–execution associations. However, we have shown that when at least three agents are trained together in a round-robin fashion, the reward rule for non-overlapping speech that we implemented causes the connection weights in each agent to converge to similar values. These connection weights are the factor in determining the action to be executed in response to an observed action. If these matrices are similar, the agents have developed similar observation–execution associations. The result is that the group of agents develops a common set of action generation rules that, in this simplified system, facilitates the coordination of speaking and non-speaking actions.

It may seem surprising that the turn balance is so smooth when one agent's motivation is increased beyond 0.5 (Figure 19.3). Because of the non-linear activation function of the neurons in the competitive choice layer, one might expect "antisocial" behavior—a non-linear response with total conversational domination once the agitation is raised above 0.5. The answer is that noise in the parallel planning layer can occasionally cause an action to win that is not the most desirable. The probability of the noise being great enough to overwhelm the desirability of speaking decreases as the desirability of speaking increases. This change in probability is linear, which yields the linear relationship observed in Figure 19.3 between agitation and conversational balance, when averaged over many trials. This is of course, assuming that the variance of the noise remains constant.

One modification made to ACQ to accommodate the turn-taking task was the exclusion of the concept of executability. Executability is based on affordances for action that are present in the environment. Gibson (1966) coined the term affordances to refer to directly perceivable opportunities for action. While the actions that Gibson considered where mainly locomotive, more recent research has investigated the nature of affordances for manual actions such as grasping (Fagg and Arbib 1998; Gentilucci 2002). It remains to be seen if there is a useful analogous concept for turn-taking actions. However, as much of turn-taking behaviors involve reactive processes, it may be useful to think of turn-taking actions being triggered by *social affordances*—that is, opportunities for social, communicative actions that are perceived "directly" (in the Gibsonian sense) with fairly minimal mental processing involved.

Our model contains multiple speaking and non-speaking actions, which, due to simplifications necessary at this initial stage, are not assigned particular meaning apart from the speaking / non-speaking distinction. Because of this, it is likely that the system would have worked with fewer actions. We expect that the need for multiple speaking and non-speaking actions for negotiating turns will emerge as the system becomes more complex and includes content generation and interpretation modules.

19.7 Conclusion and future work

This integration effort has shown promising results in extending a cognitive model of turn taking with more detailed neural modules that map to regions of the brain. We have taken the first steps towards integrating two models, YTTM (Thórisson 1996),

a cognitive model of multimodal human turn taking, and ACQ (Bonaiuto and Arbib, unpublished), a neural model of action selection. The resulting Hybrid Model is grounded in both psychological and biological research: YTTM has been implemented in virtual agents and shown to produce dynamic, human-like turn-taking behaviors in real-time through coordinated perception and generation of behaviors spanning multiple modes. The learning mechanism used by ACQ, TD learning, has been related to the basal ganglia; specifically, the TD error used to adapt the weights of the connections from the action recognition neurons to the parallel planning layer has been identified with the dopamine signal in the midbrain dopaminergic system (Schultz 1998). Winner-take-all (WTA) networks based on center-surround connectivity, like those used in ACQ, have been implicated in models of the basal ganglia (Gurney *et al.* 2001) and of networks of interacting cortical areas in reaching (Cisek 2005), and imitation (Erlhagen *et al.* 2006).

While not obvious at the outset, the two models were found to be particularly well matched for integration, primarily due to both of them having been built using compatible modular methodologies. Even less obvious was the relative success of the integration: The resulting Hybrid Model is able to evolve reactive mechanisms for turn taking, as experiments I to IV show, provided the right kind of training, motivational signal, and parameter settings. The results provide insight into how brains may accomplish cooperative communicative interaction, and suggest research directions that could lead to a more comprehensive model of turn taking in multimodal dialog. While the initial model implemented and tested here is fairly primitive, only incorporating parts from the full version of each model, the Hybrid Model presents a parameterization of turn taking in an easily extensible framework.

At the high level, the model can be extended by implementing more modules from YTTM in a neural fashion. On the lower level, elements from current work on ACQ, such as hierarchical action organization, can be included to create increasingly realistic models. Current work on ACQ involves extending the model to include the same functionality in a more detailed account of corticostriatal projections and processing within the basal ganglia. We can thus expect further biologically plausible extensions to be applied to our Hybrid Model in the near future. The implementation of YTTM as a more detailed system of neural modules lays a roadmap for the neural investigation of turn-taking mechanisms that might not have been as clearly articulated in the absence of such a high-level computational approach.

Another obvious expansion point is temporal awareness: Currently the agent selects its action based on the recognized action executed by the other agent in the previous time step. It would be better to let this decision be based on the recent history of recognized actions executed by the other agent. This would require a short-term memory, which could be based on the short-term memory module of ACQ's successor, which is currently in development, hierarchical ACQ (hACQ). To enable comparison to actual turn-taking data we will need to provide each non-speaking action and intonation with the natural constraints that each of the modes provides, that is, the models need to evolve human-like usage of intonation, gaze, and gesture. A challenge for the integration will be additional mechanisms for adding a neurally plausible version of YTTM's

timing and hierarchical action structure in ACQ. It also remains to be explored whether the temporal control scheme in YTTM maps onto such a mechanism in a convincing way.

Other elements of hACQ are candidates for further expansion of the model, especially its hierarchical composition of action programs. Theoretical constructs from the YTTM can be used to further expand the model, including parallel execution of non-speech actions. How this would be implemented in hACQ is an interesting research question that remains to be answered.

Other obvious expansion points include more varied driving goals, for example how would one model an agent that abused the implicit cooperation rule of not interrupting? How would the model behave given a goal of trying to interrupt? There are other patterns that could be proposed, based on the mental state of the listener; an example would be the speaker asking a question and the listener not knowing how or what to say in response. In this case the motivation to speak would be fairly low, yet the agent would want to indicate that he has realized that the user is expecting a reply, and thus choose to perform actions to that effect. We intend to explore these kinds of patterns by introducing some comprehension and content production, driven by high-level, dynamically changing goals.

Another clear expansion point is that the agents' non-speaking actions have not been anchored, the way speech has, in some constraints based on real turn taking. This needs to be done in order for the non-speaking actions to have some meaning in relation to real human turn taking; one way to do so would be to build a rule-based trainer that would train an agent, who could then be set up to converse with itself. Such a model would quite possibly be built to a point of being worthy of interaction with a human in real-time dialog.

Currently the agents are purely reactive and do not consider the effects of their actions on the behavior of the other agent. It could be that truly communicative actions cannot emerge without the capability to model another agent with enough fidelity to *predict* the effects of one's actions in terms of the modification of the responses of the other agent. In the same way that internal models of the world are required for skilled motor movements, internal models of other agents are required for skilled social interactions.

Acknowledgments

This work was supported in part by a Fellowship grant from Zentrum für Interdisziplinäre Forschung, a research grant from RANNÍS, Iceland, and by a Marie Curie European Reintegration Grant within the 6th European Community Framework Programme. The authors would like to thank the anonymous reviewers, Andrew Gargett for insights into Dynamic Syntax modeling, Ipke Wachsmuth for suggesting this line of research, and the ZiF fellows onboard the research train *Embodied Communication in Humans and Machines*. And finally, big thanks to Michael Arbib for extending valuable resources to this work.

References

Alstermark B, Lundberg A, Norrsell U, and Sybirska E (1981). Integration in descending motor pathways controlling the forelimb in the cat: 9. Differential behavioural defects after spinal cord lesions interrupting defined pathways from higher centres to motorneurones. *Experimental Brain Research*, **42**, 299–318.

Arbib MA (1992). Schema Theory. In SC Shapiro, ed. *The Encyclopedia of Artificial Intelligence*, 2nd edn, pp. 1427–43. NY: Wiley Interscience.

Bischoff R (2000). Towards the development of 'plug-and-play' personal robots. *1st IEEE-RAS International Conference on Humanoid Robots*. MIT, Cambridge, September 7–8.

Bonaiuto J and Arbib MA (unpublished). What Did I Just Do? A New Role for Mirror Neurons.

Botvinick MM, Braver TS, Barch DM, Carter CS, and Cohen JD (2001). Conflict monitoring and cognitive control. *Psychological Review*, **108**, 624–52.

Brooks RA (1986). Robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, **2**, 14–23. [Also MIT AI Memo 864, September 1985}.

Bryson J and Thórisson KR (2000). A three-layer design approach to character-based creative play. *Virtual Reality* [Special Issue on Intelligent Virtual Agents], **5,** 57–71.

Bullock D and Rhodes BJ (2003). Competitive queing for planning and serial performance. In MA Arbib, ed.*The Handbook of Brain Theory and Neural Networks,* 2nd edn, pp. 241–4. Cambridge, MA: A Bradford Book/ The MIT Press.

Cann R, Kempson R, and Marten L (2005). *The Dynamics of Language*. London: Academic Press.

Cisek P (2005). A computational model of reach decisions in the primate cerebral cortex. In TJ Prescott, JJ Bryson and AK Seth, eds. *Proceedings of Modeling Natural Action Selection (MNAS)*, Edinburgh, Scotland. UK: AISB Press.

Dale R and Reiter E (1996). The role of gricean maxims in the generation of referring expressions. In B Di Eugenio and NL Green, eds. *Working Notes, AAAI Spring Symposium on Computational Implicature: Computational Approaches to Interpreting and Generating Conversational Implicature*, pp. 16–20.

Duncan S (1972). Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology*, **28**, 283–92.

Duncan S and Fiske DW (1977). *Face-to-Face Interaction: Research, Methods and Theory*. Hillsdale, NJ: Erlbaum.

Erlhagen W, Mukovskiy A, and Bicho E (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, **1083**, 174–88.

Fagg A. and Arbib MA (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, **7–8**, 1277–303.

Fiddick L, Cosmides L, and Tooby J (2000). no interpretation without representation: the role of domain-specific representations and inferences in the watson selection task. *Cognition*, **77,** 1–79.

Gentilucci M (2002). Object motor representation and reaching-grasping control. *Neuropsychologia*, **40**, 1139–53.

Gibson JJ (1966). *The Senses Considered as Perceptual Systems*. Boston, MA: Houghton-Mifflin.

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

Gratch J, Young M, Aylett R, Ballin D, and Olivier P, eds (2006). *Proceedings of Intelligent Virtual Agents, 6th International Conference, IVA 2006*, Marina Del Rey, CA, USA, August 21–23, 2006. Lecture Notes in Computer Science 4133. Springer.

Grosjean F and Hirt C (1996). Using prosody to predict the end of sentences in english and french: normal and brain-damaged subjects. *Language and Cognitive Processes*, **11**, 107–34.

Guazzelli A, Corbacho FJ, Bota M, and Arbib MA (1998). Affordances, motivations, and the world graph theory. *Adaptive Behavior*, **6**, 435–71.

Gurney K, Prescott TJ, and Redgrave P (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, **84**, 401–10.

Houghton G and Hartley T (1995). Parallel models of serial behavior: Lashley revisited. *Psyche*, **2,** 25.

Iizuka H and Ikegami T (2002). Simulating turn-taking behaviours with coupled dynamic recognizers. CharIn RK Standish, MA Bedau and HA Abbass, eds. *Proceedings of Artificial Life VIII, the 8th International Conference on the Simulation and Synthesis of Living Systems* Char, pp. 142–5. Cambridge, MA: MIT Press.

Iizuka H and Ikegami T (2004). Adaptability and diversity in simulated turntaking behavior. *Artificial Life*, **10**, 361–78.

Lemon O, Bracy A, Gruenstein A, and Peters S (2001). Information states in a multi-modal dialogue system for human-robot conversation. In P Kuhnlein, H Rieser and H Zeevat, eds. *Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue (BI-DIALOG 2001), Bielefeld, Germany*, pp. 1–16..

Leβmann N, Kranstedt A, and Wachsmuth I (2004). Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max. In C Pelachaud, K R Thorisson and Z Ruttkay, eds. *Proceedings of the Workshop Embodied Conversational Agents: Balanced Perception and Action, 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS04),* New York, August 19–23. ACM Press.

Maxwell BA, Meeden LA, Addo NS, Dickson P, Fairfield N, Johnson N, Jones EG, Kim S, Malla P, Murphy M, Rutter B, and Silk E (2001). REAPER: A reflexive architecture for perceptive agents. *AI Magazine*, **22**, 53–66.

O'Connell DC, Kowal S, and Kaltenbacher E (1990). Turn-taking: a critical analysis of the research tradition. *Journal of Psycholinguistic Research*, **19**, 345–73.

Sacks H, Schegloff EA, and Jefferson GA (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, **50**, 696–735.

Sato R, Higashinaka R, Tamoto M, Nakano M, and Aikawa K (2002). Learning decision trees to determine turn-taking by spoken dialogue. *Proceedings ICSLP-02*, pp. 861–4.

Schlangen D (2006). From reaction to prediction: experiments with computational models of turn-taking. *INTERSPEECH-2006*, paper 1200-Wed3WeS.3, September 17–21, Pittsburg, Pensylvania. Kluwer Academic.

Schultz W (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, **80**, 1–27.

Sutton RS and Barto AG (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Thórisson KR (1993). Dialogue control in social interface agents. *InterCHI Adjunct Proceedings*, Amsterdam, Holland, April 24–29, pp. 139–40. New York: ACM Press.

Thórisson KR (1996). *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. Ph D Thesis, The Media Laboratory, Massachusetts Institute of Technology.

Thórisson KR (1997). Layered modular action control for communicative humanoids. In N M Thalmann and D Thalmann, eds. *Computer Animation '97*, Geneva, Switzerland, June 4–7, pp. 134–43. Los Alamitos, California: IEEE Computer Society Press.

Thórisson KR (1998). Real-time decision making in face-to-face communication. *Proceedings of the Second International Conference on Autonomous Agents*, Minneapolis, Minnesota, May 11–13, pp. 16–23.

Thórisson KR (1999). A mind model for multimodal communicative creatures and humanoids. *International Journal of Applied Artificial Intelligence*, **13**, 449–86.

Thórisson KR (2002). Natural turn-taking needs no manual: a computational model, from perception to action. In B Granstr^m, D House, I Karlsson, eds. *Multimodality in Language and Speech Systems*, pp. 173–207. Dodrecht, The Netherlands: Kluwer Academic Publishers.

Thórisson KR, Benko H, Arnold A, Abramov D, Maskey S, and Vaseekaran A (2004). Constructionist design methodology for interactive intelligences. *A.I. Magazin*e, **25**, 77–90. Menlo Park, CA: American Association for Artificial Intelligence.

Wilson M and Wilson TP (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, **12**, 957–68.

Yngve VH (1970). On getting a word in edgewise. *Sixth Regional Meeting, Chicago Linguistics Society*, pp. 567–78.