A Multiparty Multimodal Architecture for Realtime Turntaking

Kristinn R. Thórisson[1,2] Olafur Gislason[1]
Gudny Ragna Jonsdottir[2] Hrafn Th. Thorisson[1]

[1]Center for Analysis & Design of Intelligent Agents, Reykjavik University

[2]Icelandic Institute for Intelligent Machines

Menntavegur 1, 101 Reykjavik, Iceland

{thorisson,olafurgi,gudny04,hrafnt05}@ru.is

Abstract. Many dialogue systems have been built over the years that address some subset of the many complex factors that shape the behavior of participants in a face-to-face conversation. The Ymir Turntaking Model (YTTM) is a broad computational model of conversational skills that has been in development for over a decade, continuously growing in the number of factors it addresses. In past work we have shown how it addresses realtime dialogue, communicative gesture, perception of turntaking signals (e.g. prosody, gaze, manual gesture), dialogue planning, learning of multimodal turn signals, and dynamic adaptation to human speaking style. The architectural principles of the YTTM prescribe smaller architectural granularity than most other models, and its principles allow non-destructive additive expansion. In this paper we show how the YTTM accommodates *multi-party dialogue*. The extension has been implemented in a virtual environment; we present data for up to 12 simulated participants participating in realtime cooperative dialogue. The system includes dynamically adjustable parameters for impatience, willingness to give turn and eagerness to speak.

Key words: Turntaking, Dialogue, Realtime, Multiparty, Multimodal, Architecture, YTTM

1 Introduction

Traditionally, research on turntaking in natural dialogue has focused on surface phenomena, observable patterns of behavior such as speech generation, turntaking, interruptions, prosody, gaze, gestures, head and body movements, and so on. Many researchers have tried to characterize the complex interplay of *movement* and *action* in the "simplest possible way", searching for a set of rules or a "main element" or key unit with general explanatory powers (c.f. [1] [2]). In contrast to a focus on behavior, we have been addressing the many control processes underlying dialogue behavior – the *architecture of dialogue cognition*, and adopted a systems focus. Our work on turntaking extends well over a decade, and the results have been demonstrated in the Ymir Turntaking Model

(YTTM) [3] [4], which by now has been implemented in numerous interactive systems (c.f. [5] [6] [7]). Over the past decade we have expanded the YTTM in various ways. Throughout these enhancements we have preserved the YTTM's architectural principles, this being perhaps the most significant distinguishing feature of our approach: Extensions of the YTTM have been additive, meaning that prior functions included in – and explained by – the model are preserved as it gains generality. By now the model addresses a relatively broad set of features, including issues of coordinated multimodal planning and execution, perceptual organization and prioritization, and even learning.

In this paper we describe an extension of the YTTM for *multiparty turntaking*, adding to the model's existing functionalities the ability to model multiple speakers engaged in cooperative dialogue. In line with prior expansions, this particular one has not changed any of the framework's operating principles. We present data showing the behavior of the system with up to 12 simulated agents interacting in "polite" (cooperative) dialogue. While data presented here is limited to simulated scenarios; elsewhere we have tested the YTTM in realtime dialogue with human users with good results [5] [8], and evaluated its performance in comparison to human-human dialogue [9]. Although this does not replace eventual testing of the new multiparty features with human users, YTTM's prior track record in in this respect gives the simulation data presented here added weight, making it more than simply preliminary.

The paper is organized as follows: In section 2 we d iscuss theoretical framework and related work; in section 3 we discuss the YTTM, with section 4 detailing the multi-party extensions. Section 5 describes the evaluation setup and conclusions.

2 Background & Related Work

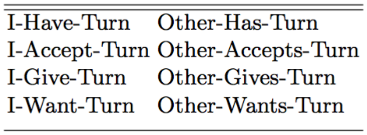
Schegloff described turntaking as that "When persons talk to each other in interaction, they ordinarily talk one at a time and one after each other" ([10], p. 207). Many have pointed out, however, that during the "live performance" of dialogue this characterization is a gross simplification at best, and at worst a rather inaccurate description of what actually happens in dialogue, as utterances frequently overlap, people interrupt themselves and others all the time, and talk on top of each other. We agree. We have argued elsewhere [11] that conversational skills belong to the class of systems that Simon referred to as *nearly-decomposable* [12], and that much of the complexity of dialogue stems from complex interaction between a set of both loosely and tightly coupled ("nearly decomposable") functions. The runtime behavior of these functions produces side effects on the observable surface phenomena that we recognize as hesitations, interruptions, and so on. A corollary can be found in much of last century's research on attention (c.f. [13] [14]), which has reached similar conclusions with regards to attentional mechanisms. Turntaking, in this view, is an indirect result of the many mechanisms at play in dialogue, in particular *their complex interaction* and effects of *limitations of realtime cognitive capabilities*. Our stance is that the best way to capture the operation of the many interacting mental functions in dialogue is to try to model dialogue as a fairly complete cognitive system, at a relatively fine level of detail.

Lessman et al. [15] describe one of the few efforts, besides YTTM, describing an explicitly cognitively motivated computational turntaking system, includng some thoughts on anticipation and the perception-action loop. They integrated turntaking mechanisms of the Max agent in a belief-desire-intention (BDI) architecture (although the BDI approach itself is not cognitively motivated). The system incorporated dialogue framework ideas from Traum and Rickel [16], who propose what they call "layers" of dialogue, although their use of the concept of layers does not seem connected to architectural concerns or cognitive principles.

The original prototype of YTTM employed close to 100 modules of various types (deciders, perceptors, etc.), as well as layers motivated by experimental data on the human perception-action cycle [3] [4] [8]. As mentioned above, YTTM has been expanded over time, at this point incorporating not only a complete set of modes (manual gesture, facial expression, head movements, eye movement, speech) for both perception and action, but also more recently realtime adaptation to human speakers' speaking style. This is a direct result of its architectural foundation, building on fine-grain modularity and a cognitively-motivated organizational structure. Much of the work of others has dismissed the YTTM's fine granularity, choosing coarser-grain approach, in line with standard software practices; one example is Traum and Rickel's work [16]; another example is Raux' [17] two-party dialogue architecture, which otherwise builds on YTTM. Coarse granularity is likely to hamper architectures' further expansion and may result in significant redesigns and changes every time the systems are improved with new functionality.

3 Cognito-Theoretical Basis of the YTTM

The YTTM [3] [4] is an agent-oriented model, motivated by a cognitive focus, taking into account top-down and bottom-up processing and making *time* a first-class citizen throughout. One of the key concepts in the YTTM is the idea of a *perception-action context* ("context" for short): an active set of perception and action processes relevant to the current situation and goal(s) of an agent. In our implementations for turntaking only one or two contexts are typically active at the same time, although parallelism is not prohibited and the context set may be changed – permanently or temporally – by attention processes that dynamically activate and deactivate the processes. Turntaking is essentially a *coupling* of these contexts – an agent synchronizes his context(s) with those of others, as inferred from behavior tracked by his perceptual processes. Contexts bear some resemblance to the approach for perception-action organization taken in behavior-based A.I., a major difference being that our approach allows dy-



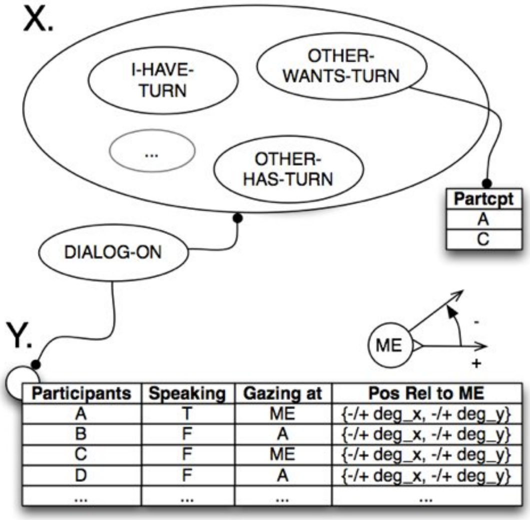
I-Have-Turn	Other-Has-Turn
I-Accept-Turn	Other-Accepts-Turn
I-Give-Turn	Other-Gives-Turn
I-Want-Turn	Other-Wants-Turn

Fig. 1. Turntaking contexts.

namic context evolution at runtime, resulting in greater flexibility and dynamic system behavior, and being more compatible with cognitive research on human attention mechanisms (c.f. [14]). Most system modules (perception, decision, behavior, etc.) in the YTTM are confined to be active only in one or at most two contexts. This is in effect a way of implementing dynamic attention control. The architecture's modularity and separation of topic knowledge and behavioral knowledge make it relatively easy to install and add increasingly complex components to the system, which forms the basis for the multiparty extension.

4 Multi-Party Extension

Our multiparty implementation includes the main principles of the above outlined functionalities, including multimodal action and perception. We have implemented eight dialogue contexts (see Figure 1), each containing the various perception and action modules, representing the disposition of the system at any point in time (see discussion of perception-action contexts, above). As in prior versions, activations of contexts are done via messages, and executed by a set of cooperating modules. In the previous model the turntaking contexts were implemented to be mutually exclusive; in the updated model the contexts *I-Want-Turn* and *Other-Wants-Turn* are allowed to be active simultaneously with other states. This was not necessary in the dyad model as *Other-Wants-Turn* would imply that "I have turn" and vice versa – in multiparty dialogue this information is not given. Further extensions include lists of participants, who is speaking, gaze perception and position perception (see Figure 2).



Partcpt
A
C

Participants	Speaking	Gazing at	Pos Rel to ME
A	T	ME	{-/+ deg_x, -/+ deg_y}
B	F	A	{-/+ deg_x, -/+ deg_y}
C	F	ME	{-/+ deg_x, -/+ deg_y}
D	F	A	{-/+ deg_x, -/+ deg_y}
...	...	...	...

Fig. 2. Extensions to implement multi-party/multimodal turntaking within the YTTM model include adding perceptions of participants, gaze, position and who is speaking (not the same as who has turn). Some information is relevant in some contexts and not others. Notice the distributed nature of the data: all data represented in the system (shown here in tables) is accessible by *any active* cognitive process that needs it.

We have implemented the complete set of contexts hypothesized for Western dialogue participants and implemented in prior versions of the YTTM, including the minimal set of perceptions and actions necessary to run the system, as proven in our prior evaluations with live human interlocutors. As before, modules are confined to be active only in certain contexts, e.g. backchannel generation only happens when someone else has turn. Multimodal deciders use information extracted from the external (perceptual) and internal environment, to synchronize the perceived and anticipated contexts, which in turn steer both perceptual and behavioral actions. All activation and de-activation of contexts is driven by perceptual events, modulated in some cases by temporally-dependent thresholds. Each agent is implemented to run in its own thread, so perceptual aliasing can happen in the system just like in human conversations. The core operating principles of the YTTM have been thoroughly demonstrated for realtime interaction with human users in [5] and [8], and our artificial agents are based on what has worked before in this respect.

Each conversation participant has an individual context model, updated with decisions from its internal deciders and input from its perception modules. Perceptions include a list of all conversation participants, who is talking, who is "looking at me" (for any given agent) and who is requesting turn at each given time. Each participant also has configuration for *urge-to-speak*; probability that another participant wants to talk (based on perceptions of their actions), the *speed at which urge-to-speak rises* (modeling a type of impatience for getting the turn), and the *yield tollerance* when someone else wants it (while he has turn). The last parameter is currently linked to the amount that an agent intends to say, so that interruptions are less likely in the middle of an utterance than at their very beginning or end. When the agent perceives that someone has the turn the agent looks at that person. When that agent gives turn (semi-explicitly) by looking at a specific person at the end of an utterance, other agents look at that person as well. For any given participant, its perception of the gaze behaviors of others determines in part whether it is possible to take turn politely. These behaviors provide only the beginning of what could be a much more elaborate set of behaviors; what is more important are the principles by which the system can easily be expanded to accommodate a much larger set of these behaviors.

5 Experimental Results

The model has been implemented in a virtual world, and we have tested it with up to 12 agents, demonstrating its scalability to relatively large groups.[1] Under a cooperation goal, participants take into consideration each other's limitations on attention, and yield if someone wants to speak, depending on the setting of their *yield* parameter. Our simulation works at the decisecond level of granularity; it should be noted that the perception of turn availability is a process that takes time, as does the decision to start speaking. Typically, overlaps only occur when two agents simultaneously decide to take turn, when their decisions are executed at a frequency above perceptual sampling rate of 10 Hz. In our scenario we have seen overlaps, but since agents are set to be very polite, typically these periods last less than 300 msecs.

[1] See video at http://www.youtube.com/watch?v=CVvNrv7K4bA

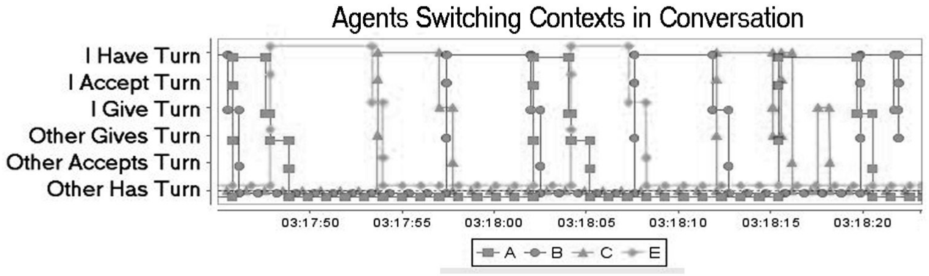


Fig. 3. The graph shows 30 seconds of conversation between 4 agents. Throughout these 30 seconds all agents' perception about who has turn is synchronized; as one agent perceives he has turn, others perceive that other has turn.

Each agent has an isolated context model and during the course of a conversation he switches contexts within the turntaking model based on his own local perceptions. If agents' perceptions agree on the proceeding of the conversation we should see a pattern of one agent having the turn at a time and others perceiving that someone else has the turn. This is precisely what we see in Figure 3. For further detail on this way of plotting context activations, and the meaning of context alignment, the reader is referred to [5].

We have evaluated the difference in behavior when all agents have a low urge to speak (5%) versus high (95%) (see Table 1). When all agents have high urge 6,6% of the total time is silence, this rises to 43,47% when urge to speak is low. These different settings do not effect average length of turn, nor overlaps.

Although further work will be needed to thoroughly evaluate the multiparty extension, earlier work shows the ability of the YTTM to address realworld, realtime dialogue [5] [8], strengthening the simulated results shown here. A thorough comparison to human group dialogue/turntaking patterns remains to be made. Future extensions include adding content generation and interpretation mechanisms, which we have demonstrated for two-party conversation [3], as well as incorporating prosody analysis demonstrated in other versions of the YTTM [5].

Table 1. Average length of turns/silences/overlaps in milliseconds.

Agents	Factor High			Factor Low		
	Portion	Average	StdDev	Portion	Average	StdDev
A	26,6%	3206	1.102	17,37%	3346	1.243
B	18,65%	3232	1.126	11,04%	2841	1.184
C	26,13%	3149	1.145	12,39%	3607	1.041
E	20,87%	4451	1.223	13,7%	3221	1.161
Overlap	1,15%	910	981	2,04%	2122	1.603
Silence	6,6%	256	159	43,47%	2505	3.832

Acknowledgments. This work was supported in part by research grants from RANNIS, Iceland. The authors wish to thank Claudio Pedica for help with the CADIA Populus environment and Erik Parr for help with the animations.

References

1. Allwood, J.: An activity based approach to pragmatics. In Black, W., Bunt, H.C., eds.: Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics, Amsterdam, John Benjamins (2000) 47–80
2. Sacks, H., Schegloff, E.A., Jefferson, G.A.: A simplest systematics for the organization of turn-taking in conversation. Language **50** (1974) 696–735
3. Thórisson, K.R.: Natural turn-taking needs no manual: Computational theory and model, from perception to action. In B. Granström, D. House, I.K., ed.: Multimodality in Language and Speech Systems, Dordrecht, The Netherlands, Kluwer Academic Publishers (2002) 173–207
4. Thórisson, K.R.: Communicative Humanoids: A Computational Model of Psycho-Social Dialogue Skills. PhD thesis, Massachusetts Institute of Technology (1996)
5. Jonsdottir, G.R., Thórisson, K.R.: Teaching computers to conduct spoken interviews: Breaking the realtime barrier with learning. In: IVA '09: Proceedings of the 9th International Conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2009) 446–459
6. Thórisson, K.R., Jonsdottir, G.R.: A granular architecture for dynamic realtime dialogue. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 131–138
7. Bonaiuto, J., Thórisson, K.R.: Towards a neurocognitive model of realtime turntaking in face-to-face dialogue. In I. Wachsmuth, M. Lenzen, G.K., ed.: Embodied Communication in Humans And Machines, U.K.: Oxford University Press. (2008)
8. Thórisson, K.R.: A mind model for communicative creatures and humanoids. International Journal of Applied Artificial Intelligence **13(4-5)** (1999) 449–486
9. Jonsson, G.K., Thórisson, K.R.: Evaluating multimodal human-robot interaction: A case study of an early humanoid prototype. Proceedings of the 6th International Conference on Methods and Techniques in Behavioral Research (2010)
10. Schegloff, E.A.: Between micro and micro: Contexts and other connections. In Alexander, J.C., Giesen, B., Munch, R., Smelser, N.J., eds.: The Micro-Macro Link, Berkeley and Los Angeles: University of California Press (1987) 207–234
11. Thórisson, K.R.: Modeling multimodal communication as a complex system. In Wachsmuth, I., Knoblich, G., eds.: Modeling Communication with Robots and Virtual Humans. Volume 4930 of Lecture Notes in Computer Science., Springer (2008) 143–168
12. Simon, H.: Near decomposability and the speed of evolution. Industrial and Corporate Change **11**(3) (2002) 587–599
13. Cavanagh, P.: Attention routines and the architecture of selection. Cognitive Neuroscience of Attention (2004) 13–28
14. Driver, J.: A selective review of selective attention research from the past century. British Journal of Psychology **92** (2001) 53–78
15. Lessmann, N., Kranstedt, A., Wachsmuth, I.: Towards a Cognitively Motivated Processing of Turn-Taking Signals for the Embodied Conversational Agent Max. In: AAMAS 2004 Workshop Proceedings: "Embodied Conversational Agents: Balanced Perception and Action". (July 2004)
16. Traum, D., Rickel, J.: Embodied agents for multi-party dialogue in immersive virtual worlds. In: AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems, New York, NY, USA, ACM (2002) 766–773
17. Raux, A., Eskenazi, M.: A multi-layer architecture for semi-synchronous event-driven dialogue management. In: ASRU, Kyoto, Japan (2007) 514–519