HUMAINE: Human Multi-Agent Immersive Negotiation Competition

Rahul R. Divekar[+]
Rensselaer Polytechnic Institute
Troy, NY, USA
divekr@rpi.edu

Hui Su[+]
Rensselaer Polytechnic Institute
IBM Research
NY, USA
huisuibmres@us.ibm.com

Jeffrey O. Kephart[+]
IBM Research
Yorktown Heights, NY, USA
kephart@us.ibm.com

Maira Gatti DeBayser
IBM Research
Brazil
mgdebayser@br.ibm.com

Melina Guerra
IBM Research
Brazil
melinag@br.ibm.com

Xiangyang Mou
Rensselaer Polytechnic Institute
Troy, NY, USA
moux4@rpi.edu

Matthew Peveler
Lisha Chen
Rensselaer Polytechnic Institute
Troy, NY, USA
{pevelm; chenl21}@rpi.edu

[+]Equal Authors

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3383001

Abstract

Competitions that directly pit software agents against one another have proven to be an effective and entertaining way to advance the state of the art in a multitude of AI domains. Less frequently, human-agent competitions have been held to gauge the relative competence of humans vs. agents, or agents vs. agents as measured indirectly by their performance against humans. We are developing a platform that supports a new type of AI competition that involves both agent-agent and human-agent interactions situated in an immersive environment. In this competition, human buyers haggle (in English) with two life-size AI agents that attempt to sell them various goods. We describe several research challenges that arise in this context, present the platform architecture and accompanying technologies, and report on early experiments with simple agents that establish feasibility and suggest that human participants enjoy the experience.

CCS Concepts

•**Human-centered computing** → **Natural language interfaces; Gestural input;** *Usability testing; Interaction design;*
•**Computing methodologies** → **Multi-agent systems;**

Introduction

For at least the past two decades, competitions that directly pit two software agents (or two *teams* of software agents)

Author Keywords:
Agent competition; Immersive environment; Multimodal dialogue; Multiparty dialogue; Negotiation; Dialogue Systems; Mixed Reality

against one another have been an effective and entertaining way to advance the state of the art in a multitude of AI domains, such as robotic soccer and agent-based electronic commerce [36]. To a lesser degree, competitions have been held to assess the relative competence of humans and agents [11] or to evaluate which agents perform best against humans [28].

There are natural scenarios in which interactions among humans *and* multiple agents are of interest. An example we draw inspiration from is an educational scenario in which students practice Mandarin Chinese language and culture through spoken role-play with embodied AI agents in an immersive environment. Initial studies of AI-assisted language education [3, 13, 17] had shown that immersion has a beneficial impact. In the Mandarin education scenario, the agents play various roles, including shopkeepers who compete with each other for the student's business [15]. As we tried to develop competent negotiation strategies that would engage the students, we realized that, with suitable extensions, our platform could be used as a basis for a new type of AI competition that blends aspects of agent-agent and human-agent interactions, and brings those interactions to life by situating them in an immersive environment (Fig. 1). The interactions also have the potential of letting users compare complex products/services in new ways[1].

With the support of members of the ANAC (Automated Negotiating Agents Competition) board (web.tuat.ac.jp/~katfuji/ANAC2019), we are preparing an international AI competition in which buyers haggle (in English) with two life-size AI agents that attempt to sell them various goods. After describing several research challenges that arise in this context and a brief literature review, we present the platform architecture and accompanying technologies. Finally, we report on early experiments with simple agents that establish feasibility and suggest that human participants find the experience enjoyable.

[1]E.g. users could interact with agents that represent different products (e.g. competitor car brands) or perspectives on the same product (e.g. style consultant, engineer).

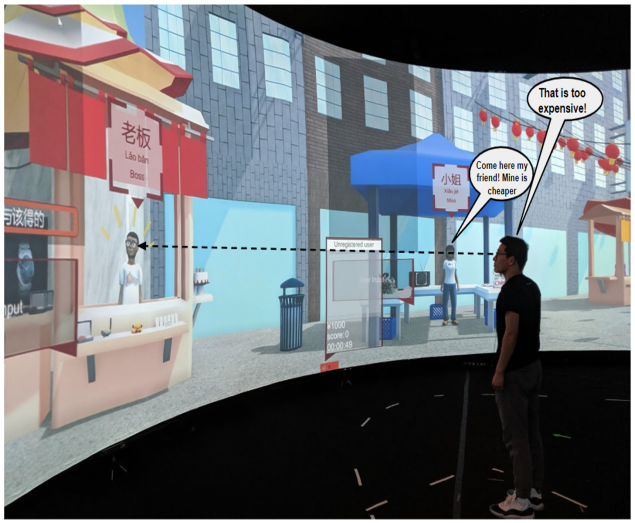

Figure 1: Multi-modal immersive environment for AI competition.

Literature Review

Two essential challenges that arise in the context of non-dyadic interactions among humans and agents include how an agent can know a) when it is being addressed and b) when it may speak.

Several authors, including [33], [38], [35], [1], [25], [5] and [31]) have sought means of determining the addressee without resorting to a wake word by means of various multimodal cues such as intonation, pitch, head-gaze, vocal energy, etc. to determine the addressee in human-kiosk, human-robot, human-human, and human-human-agent conversations. [23] [35], [30], [20] and many others have addressed this problem in the Human Robot Interaction

field using approaches such as identifying visual focus of attention or moving the robot's head to signify turns. Recent work [14] shows that a simple approach based on head pose coupled with semiotics of inferred user attention by the avatar may suffice. Encouraged by this result, we adopt a similar approach but employ a different head pose recognition algorithm [10] that copes better with the low light and high pose angle that typifies our environment (Fig. 1).

Multi-party, multi-modal conversations have been studied from various perspectives (e.g. linguistics, algorithms) and embodiments (kiosk based, etc.) [4], [24], [19], [7], [37], [2], [32]; but rarely in a multi-agent competitive, immersive setting with natural interactions. We find the recent work of Gatti de Bayser et al. [18], [12] to be most appropriate for our use case. Their approach based on deontic logic explicitly models turn-taking in conversations involving humans and multiple AI agents. It can enforce rules on the structure of the conversation without requiring individual agents to understand or implement the model themselves.

AI competitions that pit software agents against one another have a long history that includes annual events with multiple competitive leagues, such as the RoboCup robotic soccer competition (https://www.robocup.org/), the Trading Agent Competition (https://strategicreasoning.org/trading-agent-competition/, and the Automated Negotiating Agents Competition (web.tuat.ac.jp/~katfuji/ANAC2019) [26, 22, 29]. Among the smaller set of human-agent competitions in existence is the ANAC human-agent league [27], wherein humans negotiate with a single agent using a desktop-based interface that allows them to input structured text supplemented with emojis to indicate emotion. As far as we are aware, ours is the first competition that features direct simultaneous negotiations between a human and multiple competing agents. Our competition is further distinguished by occurring in an immersive environment that incorporates multi-modal interactions involving speech understanding, speech synthesis, and addressee detection based upon head pose estimation – all of which combine to support a more natural form of interaction between humans and software agents thereby attempting to provide a sense of realism, visual and social presence (defined as sense of being with another intelligent entity [6]).

[2]We use head pose because studies of similar repeated-interaction scenarios indicate that users find it more natural than using a wake word [13]

Negotiating in an Immersive Environment

The purpose of the immersive environment is to provide audio-visual immersion and presence, i.e. a feeling of being in a different place. Fig. 1 illustrates the 360-degree panoramic screen that is used to provide visual immersion; audio immersion is provided via spatial audio techniques [9, 8] that enable one to control the apparent location of sound sources. While similar physical immersive systems exist [21, 34], they tend to be used for other situational contexts such as combat simulation.

The screen depicts a virtual street scene in Shanghai (implemented in Unity Game Engine) inhabited by two street vendor avatars. Wearing a lapel microphone, the user looks at the avatar with whom they wish to speak. A central system transcribes the speech and infers the addressee using head pose information[2], and forwards this information to all of the agents. The central system maintains decorum and fairness by using a Finite State Automaton to enforce certain predetermined turn-taking rules, including ones that specify when interjections are permitted.

During each of several rounds in the competition, a human buyer starts by stating which goods they are interested in purchasing (e.g. eggs, milk, sugar, flour, chocolate) in an effort to acquire ingredients to complete a task (e.g. bake cakes). Agents may choose to respond to such requests

with offers consisting of a bundle of ingredients and a price. Such offers are rendered as synthesized speech in such a way as to appear to emanate from that agent. The human may respond with a counter-proposal, and so on until an agreement is reached. While humans may primarily direct their attention to one agent, all agents are aware of all negotiation messages exchanged in the system, and (under conditions regulated by a Moderator) they may interject with convincing counter-arguments/offers, possibly causing the human to re-direct their attention. An accompanying video[3] illustrates such interactions. Agents are evaluated quantitatively according to their total financial gain, while humans are evaluated according to a utility function that depends on the number and quality of cakes that can be assembled from their ingredients. Agents may also be judged qualitatively according to their perceived degree of engagement.

[3]Interaction Demo Video: https://youtu.be/KBV9z9fLAD0

Architecture and Technical Details

Extended from [16] which was limited to single agent, wake-word based interactions, Fig. 2 shows the overall architecture that enables the interactions described above. Each module in the diagram can reside on a different machine and can communicate using publish-subscribe software (RabbitMQ) and RESTful API.

In step 1, the *Mic* and *Cameras* detect raw input. The mic connects to a commercial cloud-based Automatic Speech Recognition (ASR) service. The machine connected to the camera processes images to detect head pose according to a method described by [10]. In step 2, the text utterance and head pose coordinates are sent to the *Attention Manager*, which infers the addressee as the agent towards which the human was looking primarily over the course of the utterance (Agent 1 in our example). In step 3, the *Attention Manager* forwards the utterance U, the speaker S, and the inferred addressee A to the *Moderator*, which updates the global conversation state and if the utterance is allowed, forwards it to the *Agent Executors*, which generate dialogue and visual actions of the avatars that embody the agents.

The architecture permits participant developers to submit independently-written *Agent Executors* that understand the messages generated by the system, compute actions (offers, acceptances, etc.) according to their strategies (for instance, bidding strategy), and speech utterances that represent those actions. The only other requirement is that the agents provide a list of possible responses ahead of time to enable the intent classifier of the *Moderator* to be trained.

Both *Agent Executors* receive the utterance, speaker and addressee, and they may choose to generate a proposed utterance. The architecture places no restrictions on the utterance or on the means by which it is produced. In our current implementation of the agents, they use an intent-entity based dialogue engine along with local conversation state variables to select a dialogue node. They use a naive negotiation strategy in which the default bidding agent's behavior simply decrease the last heard bid by a predefined amount until a predefined lower limit is reached. *Agents Executors* may propose any response they want to speak and pass it to the moderator in step 4. For example, we shall use the example of Table 1: after the user said “I want to buy tea" while looking mostly at Agent 1, the Agents 1 and 2 might propose “Yes, for $5" and “Yes, for $6", simultaneously and pass it to the *Moderator*.

The *Moderator* is a centralized controller that regulates the interaction, and protects against either of the agents unfairly (or annoyingly) hijacking the interaction. It has its own separate intent classification engine. This helps maintain a global notion of intents that may differ from that of the agents. It consults the classification engine in steps 4.1 and 4.2 to get the intent of each proposed utterance. This

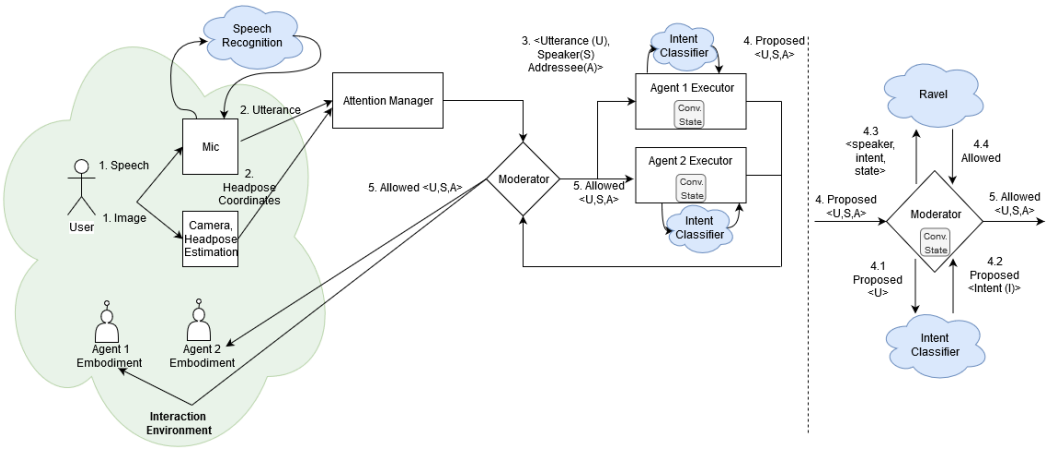

Figure 2: MMIDA: Multimodal Multiagent Immersive Dialogue Architecture

T	S	H	Utterance	Status	Rule
t1	U	A1	I want to buy tea	Broadcast	R1
t1.1	A1		Yes, for $5	Broadcast	R3
t1.1	A2		~~Yes, for $6~~	Block	R3
t1.2	A2		I can give it to you for cheaper	Broadcast	R4

Table 1: Sample Dialogue - Multi-party turn taking. T: Turn; S: Sender; H: Head pose

intent, along with global conversational state it maintains, is sent in step 4.3 to *Ravel* [18] [12] — a Finite State Automaton (FSA) that evaluates the information and decides whether the utterance is allowed. We have designed the following regulatory rules for the competition: **R1:**User is always *allowed* to reply.; **R2:** AI Agents are *prohibited* from self-responses.; **R3:** If direct addressee detected, it has the *obligation* to respond; other agents are *prohibited*.; **R4:** AI Agents are *allowed* to respond to a price pitch. Thus, in Table 1, Agent 1's pitch (t1.1, A1) passes through while that of Agent 2 is blocked (t1.1, A2). The blocking helps from agents speaking over one another. A2 will be allowed in the next turn to counter-reply A1's accepted utterance. The *Moderator*, in step 5 (Fig. 1), passes the allowed utterance on to the appropriate avatar, which is rendered using the Unity Engine and a commercial text-to-speech engine in conjunction with the spatial audio system [8]. The *Moderator* also sends the accepted utterance to both the *Agent Executors*. Note that this time the attention manager can be skipped as the addressee is the entire room. Here, the agents upon hearing each others' (or their own) bids may continue the cycle afresh by proposing a counter-pitch (seen in t1.2, A2). The counter pitches are received by the *Moderator*, at most one is allowed, and so the cycle continues until an agreement is reached. The architecture is

imaginably scalable to more participants and conversational contexts[4].

Discussion from various perspectives

Administrator's Perspective
An administrator's role is split into two parts: pre-competition preparation and in-competition facilitation. Prior to the start of the competition, she requests and collects all potential phrase variations that the agents may utter and submits them to the *Moderator* and *Ravel*. During the competition itself, she uses a web-based-backend UI to generate rounds, indicate start/end of competition to parties, generate utility functions that provide incentives for agents and human competitors and, validate the final offers[5].

Participant Developer Perspective
At the beginning of a round, the agents receive fresh utility functions from the Admin, which serve as incentives that drive their negotiation behavior. We expect that the competition will provide wide scope for research on both the strategic and psychological aspects of negotiation with humans; for example the agents may try to gain an advantage by expressing their bids in an engaging or attractive way, or by dissing other agents. As a first test of the platform, two external groups successfully created agents in a pilot test. Detailed documentation and sample code will be made publicly available well in advance of the competition.

Human Negotiator Perspective
Two in-house agents were employed with the main purpose of role-playing haggling with users to learn a foreign language. We used this as a pilot to test whether the proposed competition sufficiently engages human participants. 13 college students (6 female, 7 male) participated. Prior to the study, they were told how to direct an utterance towards an agent using head pose. They were *not* told that other agents might interject even when not addressed. We evaluated the interaction using a post-experience questionnaire.

To judge the overall experience, we asked users to rate on a Likert scale of 1-5 whether they agreed that the interaction was usable and (in a separate question) likable. The responses to usability had a mean of 4.08 ± 0.86, while the responses to likability had a mean of 4.38 ± 0.65. A one-sample test revealed that with p=0.003 and p=0.0002 we could say that the true median for usability and likability was greater than 3 (neutral). We also asked them to rate the appropriateness of agent's turn taking on a Likert scale. The mean score for responses to it was 4.36 ± 0.51. We did a One-sample Sign-test on the data and with confidence of p=0.0048 we could say that the true median was greater than 3 (neutral). Overall, we find that the interaction closely matched the users' natural expectations and thus the design appears to suffice for the competition.

Conclusion and Future Work

We have provided an overview of the architecture and technology underlying a new AI competition that we hope to hold as part of ANAC at an upcoming AI conference: IJCAI 2020. Our preliminary evaluations with simple agents and a small number of human participants indicate that the platform supports independently-programmed agents and that humans find the experience engaging. Such conversational interactions may augur a new way for businesses to advocate their wares and users to compare attributes of complex goods or services. We hope and believe that this multi-agent, multi-modal negotiation platform and the competition it supports will — as RoboCup and TAC (Trading Agent Competition) have done before it — spur new and interesting research in the realm of multi-lateral negotiation algorithms, multi-modal dialogue, and the human dynamics of such interactions.

[4] Similar turn taking rules have worked for other conversational contexts shown in [18]. More participants and agents can be accommodated with more sensors as shown in [3][39]

[5] The admin takes no role in the dialogue or turn-taking.

REFERENCES

[1] Oleg Akhtiamov, Maxim Sidorov, Alexey A Karpov, and Wolfgang Minker. 2017. Speech and Text Analysis for Multimodal Addressee Detection in Human-Human-Computer Interaction.. In *INTERSPEECH*. 2521–2525.

[2] Rieks op den Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue*.

[3] David Allen, Rahul R Divekar, Jaimie Drozdal, Lilit Balagyozyan, Shuyue Zheng, Ziyi Song, Huang Zou, Jeramey Tyler, Xiangyang Mou, Rui Zhao, and others. 2019. The Rensselaer Mandarin Project—A Cognitive and Immersive Language Learning Environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9845–9846.

[4] Sean Andrist, Dan Bohus, Bilge Mutlu, and David Schlangen. 2016. Turn-taking and coordination in human-machine interaction. *AI Magazine* 37, 4 (2016), 5–6.

[5] Naoya Baba, Hung-Hsuan Huang, and Yukiko I Nakano. 2012. Addressee identification for human-human-agent multiparty conversations in different proxemics. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM, 6.

[6] Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 456–480.

[7] Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*. Association for Computational Linguistics, 98–109.

[8] Samuel Chabot and Jonas Braasch. 2017. An Immersive Virtual Environment for Congruent Audio-Visual Spatialized Data Sonifications. Georgia Institute of Technology.

[9] Samuel Chabot, Wendy Lee, Rebecca Elder, and Jonas Braasch. 2018. Using a multimodal immersive environment to investigate perceptions in augmented virtual reality systems. Georgia Institute of Technology.

[10] Lisha Chen, Hui Su, and Qiang Ji. 2019. Face Alignment with Kernel Density Deep Neural Network. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[11] Rajarshi Das, James E Hanson, Jeffrey O Kephart, and Gerald Tesauro. 2001. Agent-human interactions in the continuous double auction. *IJCAI* (2001), 1169–1178.

[12] M. Gatti de Bayser, C. Pinhanez, H. Candello, M. A. Vasconcelos, M. Pichiliani, M. Alberio Guerra, P. Cavalin, , and R. Souza. 2018. Ravel: a MAS Orchestration Platform for Human-Chatbots Conversations. In *The 6th International Workshop on Engineering Multi-Agent Systems (EMAS @ AAMAS 2018)*. Stockholm, Sweden.

[13] Rahul R Divekar, Jaimie Drozdal, Yalun Zhou, Ziyi Song, David Allen, Robert Rouhani, Rui Zhao, Shuyue Zheng, Lilit Balagyozyan, and Hui Su. 2018. Interaction challenges in ai equipped environments built to teach foreign languages through dialogue and task-completion. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 597–609.

[14] Rahul R Divekar, Jeffrey O Kephart, Xiangyang Mou, Lisha Chen, and Hui Su. 2019a. You talkin' to me? - A practical attention-aware embodied agent. In *Human-Computer Interaction – INTERACT 2019*.

[15] Rahul R Divekar, Xiangyang Mou, Lisha Chen, Maira Gatti de Bayser, Melina Alberio Guerra, and Hui Su. 2019b. Embodied Conversational AI Agents in a Multi-modal Multi-agent Competitive Dialogue. IJCAI.

[16] Rahul R Divekar, Matthew Peveler, Robert Rouhani, Rui Zhao, Jeffrey O Kephart, David Allen, Kang Wang, Qiang Ji, and Hui Su. 2018a. Cira: An architecture for building configurable immersive smart-rooms. In *Proceedings of SAI Intelligent Systems Conference*. Springer, 76–95.

[17] Rahul R Divekar, Yalun Zhou, David Allen, Jaimie Drozdal, and Hui Su. 2018b. Building Human-Scale Intelligent Immersive Spaces for Foreign Language Learning. *iLRN 2018 Montana* (2018), 94.

[18] M. Gatti de Bayser, M. Alberio Guerra, P. Cavalin, and C. Pinhanez. 2018. Specifying and Implementing Multi-Party Conversation Rules with Finite-State-Automata. In *Proc. of the AAAI Workshop On Reasoning and Learning for Human-Machine Dialogues, DeepDial'18*.

[19] Agustín Gravano and Julia Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 253–261.

[20] Erdan Gu and Norman I Badler. 2006. Visual Attention and Eye Gaze During Multiparty Conversations with Distractions. *Lecture Notes in Computer Science* 4133, 4133 (2006), 193–204.

[21] Randall W Hill Jr, Jonathan Gratch, Stacy Marsella, Jeff Rickel, William R Swartout, and David R Traum. 2003. Virtual Humans in the Mission Rehearsal Exercise System. *Ki* 17, 4 (2003), 5.

[22] Dave de Jonge, Tim Baarslag, Reyhan Aydoğan, Catholijn Jonker, Katsuhide Fujita, and Takayuki Ito. 2019. The Challenge of Negotiation in the Game of Diplomacy. In *Agreement Technologies 2018, Revised Selected Papers*, Marin Lujak (Ed.). Springer International Publishing, Cham, 100–114.

[23] M. Katzenmaier. 2004. *Identifying the addressee in human-human-robot interactions based on head pose and speech*. Ph.D. Dissertation. Carnegie Mellon University, USA and University of Karlsruhe TH, Germany.

[24] Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. 2016. Reinforcement Learning for Turn-Taking Management in Incremental Spoken Dialogue Systems.. In *IJCAI*. 2831–2837.

[25] Thao Le Minh, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. 2018. Deep Learning Based Multi-modal Addressee Recognition in Visual Scenes with Utterances. *IJCAI 2018* (2018), 1546–1553. DOI: http://dx.doi.org/10.24963/ijcai.2018/214

[26] Raz Lin, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M. Jonker. 2014. Genius: An Integrated Environment for Supporting the Design of Generic Automated Negotiators. *Computational Intelligence* 30, 1 (2014), 48–70. DOI:http://dx.doi.org/10.1111/j.1467-8640.2012.00463.x

[27] Johnathan Mell and Jonathan Gratch. 2016. IAGO: interactive arbitration guide online. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1510–1512.

[28] Johnathan Mell, Jonathan Gratch, Tim Baarslag, Reyhan Aydoğran, and Catholijn M Jonker. 2018. Results of the first annual human-agent league of the automated negotiating agents competition. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 23–28.

[29] Yasser Mohammad, Enrique Areyan Viqueira, Nahum Alvarez Ayerza, Amy Greenwald, Shinji Nakadai, and Satoshi Morinaga. Supply Chain Management World: A benchmark environment for situated negotiations. (????).

[30] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems* 1, 2 (2012), 1–33. DOI: http://dx.doi.org/10.1145/2070719.2070725

[31] Atta Norouzian, Bogdan Mazoure, Dermot Connolly, and Daniel Willett. 2019. Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed. *arXiv preprint arXiv:1902.00570* (2019).

[32] Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky. 2013. Situated multiparty interaction between humans and agents. In *International Conference on Human-Computer Interaction*. Springer, 107–116.

[33] Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[34] Jeff Rickel, Stacy Marsella, Jonathan Gratch, Randall Hill, David Traum, and William Swartout. 2002. Toward a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems* 17, 4 (2002), 32–38.

[35] Samira Sheikhi and Jean Marc Odobez. 2015. Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters* 66 (2015), 81–90. DOI: http://dx.doi.org/10.1016/j.patrec.2014.10.002

[36] Peter Stone. 2003. Multiagent Competitions and Research: Lessons from RoboCup and TAC. In *RoboCup 2002: Robot Soccer World Cup VI*, Gal A. Kaminka, Pedro U. Lima, and Raúl Rojas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 224–237.

[37] David Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*. ACM, 766–773.

[38] TJ Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia* 17, 9 (2015), 1550–1561.

[39] Rui Zhao, Kang Wang, Rahul Divekar, Robert Rouhani, Hui Su, and Qiang Ji. 2018. An immersive system with multi-modal human-computer interaction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 517–524.