Speaker Hand-Offs in Collaborative Human-Agent Oral Presentations

ABSTRACT

Prior studies of public speaking behavior have demonstrated that public speaking anxiety monotonically decreases with the number of co-presenters giving an oral presentation and increases with the size of the audience. However, speaker “hand off” behavior—the verbal and nonverbal cues used to transition from one speaker to another—and its effect on speaker anxiety and presentation quality has not been systematically studied. In this work we report on two empirical studies of speaker hand-off behavior used during human co-presentations. We find that the cues used for hand-offs during prepared and rehearsed presentations differ significantly from the cues observed in face-to-face conversational turn-taking. We describe two systems that leverage automatic recognition of these verbal and nonverbal cues to drive hand-offs during co-presentations with a life-sized virtual agent.

CCS CONCEPTS

• **Human-Computer Interaction** → Usability and acceptability, Conversational agents

KEYWORDS

Human-centered computing, Natural language interfaces, Turn-taking

1 INTRODUCTION

Prepared oral presentations are an unavoidable and recurrent event in most professions. Unfortunately, the typical quality of these presentations is poor, and at least a third of the population suffers from public speaking anxiety. One possible solution to both of these problems is co-presentation, in which two or more speakers share the stage and trade-off giving parts of the presentation. Social impact theory predicts that public speaking anxiety decreases with the number of co-presenters on the stage and increases with the size of the audience; a result that has been empirically validated. In one study, 60 participants were recruited and asked to imagine themselves in one of 72 different performance scenarios. In each of these scenarios, participants were shown images of the co-presenters and audience they should imagine themselves performing in front of. These images depicted different possible audiences and co-presenters varied in both size and social status. Their results showed that as the size of the audience increased, the participants imagined tension grew. However, as the number of co-presenters increased, the imagined tension decreased logarithmically. To further explore this finding, a second experiment was conducted in which 48 student performers were asked to fill out a questionnaire prior to giving a live performance. Results of this experiment matched the previous findings, in which presenters performing with as few as one co-presenter experienced an exponential decrease in reported nervousness.

These results have motivated the development of virtual agents that serve as co-presenters in the delivery of oral presentations [1, 2]. While these have been found to be effective at reducing public speaking anxiety and increasing audience engagement and perceived presentation quality, real-time interaction with the agent was driven by a remote control that speakers found challenging to use. This motivated a study of the verbal and nonverbal cues that human co-presenters use when giving an oral presentation and how these differ from turn-taking cues used in face-to-face conversation. We further wanted to explore whether automated detection of these cues could be used to provide a more intuitive speaker hand-off mechanism when co-presenting with a virtual agent.

In the remainder of this paper we first review related work before presenting the results from two observational studies of hand-off cues in human co-presentations and how these differ from conversational turn-taking. We then present two different systems for supporting automated speaker hand-off with a virtual agent, using cues observed in our observational studies, along with evaluations of each method, before concluding.

2 RELATED WORK

In this section we briefly review related work on technologies to support public speaking that leverage conversational behavior, and for human-agent conversational turn-taking.

2.1 Presentation Technologies

Several systems have also been developed to enable speakers to control presentation media using gestural interfaces. Baudel et al. [3] developed Charade, a system that enabled presenters to interact with their presentation using free-hand gestures. Using signals from a DataGlove, the system was capable of recognizing 16 gestural commands designed for slide navigation and interaction with slide visuals. For example, the presenter could move the hand from left to right to advance to the next slide, and could highlight part of the presentation screen by pointing with the index finger and circling the target area. Results of a user evaluations study showed that the system achieved high accuracy,

IVA ’18, November 5-8, 2018, Sydney, NSW, Australia
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6013-5/18/11$15.00
https://doi.prg/10.1145/3267851.3267904

ranging from 72-84% for first-time users and 90-98% for trained users. While these results are promising, this approach requires the user to wear a DataGlove, which can be cumbersome. More recently, researchers have used less invasive methods to capture gestural information from speakers using motion sensing devices such as Kinect. Sommool et al., developed an interactive framework for e-learning that allowed teachers to control presentations through a combination of verbal and gestural commands [4]. In their work, vocal commands were primarily used for controlling general system functionality such as turning the slideshows on or off, while gestural commands were primarily used for actions that require more precise control, such as slide navigation. In an evaluation study with 20 participants, the system was rated as being satisfying to use and more functional than traditional teaching tools. However, the study did not evaluate audiences' perceptions of presentation quality with the system.

In addition to the DynamicDuo system described in Section 4, other researchers have explored using virtual agents to deliver oral presentations. Noma's 3D virtual presenter system enabled users to annotate a presenter's speech text with various gesture commands, which could be performed by a 3D animated computer character capable of non-verbal behaviors and synthesized speech [5]. Additional interaction options could also be programmed into the system using a menu-based scripting template. Nijholt et al. explored the creation of an embodied virtual presenter agent for use in a virtual meeting room [6]. Similar to Noma's work, a virtual presenter system was developed to parse manually annotated presentations into animation scripts for a 3D character. Unlike the previous work however, this system was designed for real-time use in a virtual meeting environment. The system could capture audience motions via cameras placed in a meeting room, thereby allowing for a more realistic simulation of audience members and the person controlling the virtual presenter. No evaluation results were reported for either of these systems.

2.2 Conversational Turn-taking with Virtual Agents

Conversational turn-taking is a complex multi-modal process in which myriad cues, gaze, speech, hand gesture, and prosody, are used to coordinate the behavior of speakers [7-10] so that they do not speak at the same time. One of the general hypotheses from this work is that the overall strength of a turn-taking signal is a function of the number of cues involved [11].

Several computational models have been developed to enable humans to engage in conversational turn-taking with dialogue systems, virtual agents and robots using automated recognition and production of these cues. Some researchers have explored specific cues associated with particular aspects of turn-taking, including identification of the end of a user's turn [12], or when an agent should take or give the turn [13-16]. Others have developed complete computational models that enable users to engage virtual agents in face-to-face conversation. Early examples, including Gandalf [17], Rea [18], and the Virtual Rapport agent [19] used hand-coded rules to integrate multimodal turn-taking cues to determine when the agent should take or give the turn. More recent work has explored the development of agents that dynamically learn turn-taking cues in real-time while interacting with users [20], or more advanced action selection models such as timed petri nets [21]. Other researchers have modeled turn-taking as a continuous dynamic control problem to account for interruptions, overlaps, and silences [22].

3 HAND-OFFS IN HUMAN CO-PRESENTATIONS

To understand the verbal and nonverbal behavior that human co-presenters use during hand-offs, we review two studies of oral presentations.

3.1 Hand-Offs During TED Talks

We previously conducted an analysis of TED talks (www.ted.com) to understand co-presentation behavior in exemplary oral presentations [BLIND]. Of the 1,732 TED talks retrieved, only 34 (1.9%) were given by two presenters. From our analysis we identified several interaction formats, including: **Iterative turn-taking** (47% of talks), in which presenters take turns giving parts of the presentation; **Single turn** (27%) in which each presenter spoke exactly once; **Dialogue** (3%) featuring a staged dialogue for the entire presentation; **Interview** (3%) in which an interviewer asked each of the co-presenters a question as a prompt for the next part of their talk; and **Debate** (3%) in which a moderator introduced the co-presenters, gave them each a fixed time to make an argument, then opened the floor for interaction.

We also analyzed the verbal and non-verbal behavior used by co-presenters, particularly in the iterative turn-taking and single turn formats. Explicit verbal turn transitions (e.g., "Sean's going to tell you…") occurred in only 30% of talks, and only once or twice in each of these. Nonverbal behavior was obscured in 29% of the turn transitions (with the video showing a presentation slide). Of the visible transitions, the current presenter gazed at the next speaker when it was their turn to speak 36% of the time, and gestured at them 6% of the time. Gaze transitions appeared to most frequently represent listener behavior – in which the current speaker is passively attending to the next speaker once they start – and less frequently appear to be proactive signals to the next speaker that it was their turn.

3.2 Hand-Offs in a Design Study

We conducted a secondary analysis of data collected from a workshop in which dyads of participants each co-delivered a 5-minute presentation of a pre-selected conference paper [BLIND]. Each dyad received their assigned paper two days prior to the workshop to prepare and rehearse. We recruited 12 students and professionals (5 male, 7 female, ages 20-54, mean 30) with varying levels of presentation experience and backgrounds in computer science, communication, and life sciences. Of the 12 participants, 3 were categorized as high competence public

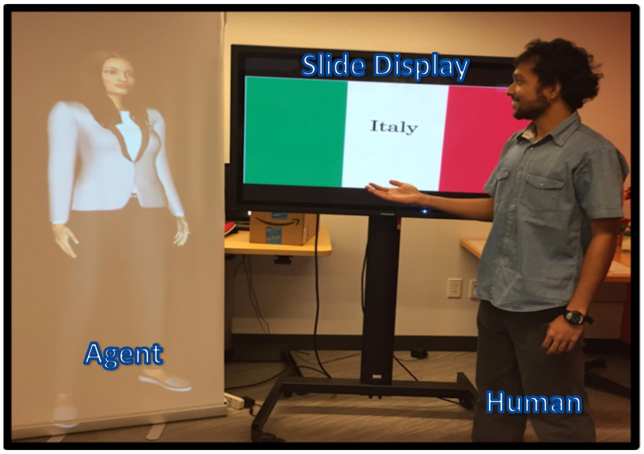

Figure 1: DynamicDuo Co-Presentation System

speakers, 2 as low competence public speakers, and 7 as moderate competence according to the Self-Perceived Communication Competence Scale [23]. Co-presentations were coded for verbal and nonverbal behavior during speaker hand-offs.

The 6 presentations had between 2 and 15 (median 4) hand-offs. Hand-offs occurred during Powerpoint slide changes 59% of the time, accompanied by gaze cues and hand gestures towards the next speaker 44% and 21% of the time, and speaker pauses 41% of the time, verbal transitions 22% of the time.

3.3. Differences between Hand-Offs and Conversational Turn-Taking

There are several significant differences between the hand-off behavior observed in our studies and face-to-face conversational turn-taking behavior from the literature. These differences are likely due to that fact that oral presentations, such as those in our studies above, are highly rehearsed performances in which one of the objectives is to avoid using coordination behaviors to actually negotiate the floor in real time in front of the audience, which could be perceived as a lack of coordination and rehearsal. We observed that over 50% of hand-offs in both corpora occurred without any observable nonverbal behavior. Of the nonverbal hand-off behavior observed, gazing at the next speaker was the most commonly observed, followed by speaker pauses and hand gestures.

4 THE DYNAMIC DUO SYSTEM

In this work, we use the DynamicDuo System [1] as our experimental testbed for exploring human-agent hand-offs. Implemented as an add-in to PowerPoint, DynamicDuo provides a life-sized animated, virtual co-presenter agent for public speaking support (Fig. 1).

The virtual presenter, Angela, is an animated human-like character developed in Unity. Angela speaks using synthetic speech and is capable of displaying a variety of nonverbal behaviors, including facial expressions of affect (smile, neutral, concern), eyebrow movement, directional gazes, head nods, posture shifts, and contrastive, beat, and deictic gestures (e.g. pointing to a slide). The majority of the agent's nonverbal behaviors are automatically generated using BEAT [24].

The baseline DynamicDuo system uses a wireless remote-control device ("clicker") to advance slides and to give the speaking turn to the agent or stop the agent's speech. Although feedback on DynamicDuo by evaluation study participants was positive, presenters did feel that use of the clicker during presentations was problematic (e.g., *"It would be cool if she could know when it is her turn to talk without me having to click the button", "not hitting the remote on time might create awkward silence"*). The following sections describe two systems that automate recognition of speaker hand-off cues.

5 HAND-OFFS USING SPEECH CUES

Based on our observational studies, we first explored the use of speech and prosody as co-presenter hand-off cues. We used pauses in the human presenter's speech as the primary hand-off cue since the most frequently used interaction format was a "seamless" transition without any accompanying verbal or nonverbal behavior. Thus, the presenter could interact with the agent in four different ways:

Short Pause at the End of Slide: Presenter could pause after they finish delivery of all sections in the Main Points segment of a slide to trigger the agent to deliver the Transition segment.

Long Pause: When the presenter paused for more than three seconds, the agent delivers the next segment of the presentation.

Verbal Turn-Taking: The presenter could ask the agent to continue delivering the next section in the presentation through verbal commands such as: "Angela, would you please continue?", "Please continue", or "Angela will present the next section".

Stop Command: The presenter could utter any words while the agent was talking to explicitly and immediately interrupt it.

5.1 Implementation

The system uses speech recognition to detect a list of key phrases used for hand-offs. The open source Pocketsphinx automated speech recognition (ASR) system was used to spot key phrases and activity in the human presenter's speech input [25]. The speech signal was also processed to detect pauses as regions in the audio signal in which no voice activity was detected. The system ignored speech segments shorter than 50 milliseconds. Two pause lengths, short and long, were detected. A short pause length was defined as at least 80 milliseconds and a long pause was at least 3 seconds long. When the agent was speaking, any voiced activity from the human presenter longer than 80 milliseconds was considered an interrupt signal.

5.2 Evaluation

A two-session within-subjects experimental study was designed to evaluate the speech-based hand-off detection system. Fifteen participants were recruited to co-present with the agent using the speech-based interface, using the original DynamicDuo's clicker based interface as a comparison condition. Approximately 50% of the presentation was given by the agent. Each participant presented two comparable pre-made presentations on topics of tigers and lions, assigned to each treatment in a counterbalanced order. Each deck had 6 slides with detailed speaking notes containing 6-10 sentences each. For each presentation, participants had 10 minutes to prepare, 5 minutes to rehearse with the agent and then another 5 minutes to deliver their final video-recorded presentations.

5.2.1 Study Measures

The self-report measures used to assess system acceptance were:

- Communication Competence: Assessed at intake using the Self-Perceived Communication Competence Scale [23].
- State Anxiety: Assessed prior to any presentation given using the State Anxiety questionnaire [26].
- Speaker Confidence: Assessed at intake and after any presentation using the Personal Report of Confidence as a Speaker questionnaire [27].
- Virtual Co-presenter Rating: Assessed after any human-agent presentation using 6-question,7-point scale (Table 1).
- Semi-structured interview: Given at the end of each session to assess general impressions, preferences and concerns.

5.2.2 Results

ASR accuracy was assessed post-study by manually annotating use of speech commands in the recorded presentations, finding relatively high accuracy rates of up to 71.5% (SD=24.6%).

Table 1 shows results of agent ratings of the two conditions. There were no significant differences between study conditions on any self-report measures. In post-study interviews, participants mentioned that they liked the idea of interacting with the agent using speech during their presentation, e.g. "*it was interesting to have that direct communication with that other person [Agent]*" [P1]. However, participants worried about the reliability of the system and did not trust the pause-based interaction, fearing that: "*the agent would jump in when I did not want it to.*" [P3]. Concerns were also brought up about the specific verbal prompts required to control the agent, suggesting need for personalization.

Table 1: Agent Ratings for Speech vs. Remote Control

Ratings of Co-presenter: 1 - Not at all; 7 - Very Much	Speech Mean (SD)	Remote Mean (SD)
How *satisfied* are you with…?	5.53 (1.50)	5.27 (1.49)
How much do you *like*…?	5.6 (1.12)	4.87 (1.64)
How much do you feel you	5.2 (1.52)	5.6 (1.50)
How *helpful* was…?	5.67 (1.23)	4.73 (2.05)
How much would you like to *give future presentations* with…?	5.73 (1.49)	4.87 (1.73)
How *easy* was it to use…?	3.47 (2.10)	4.2 (1.93)

5.2.3 Discussion

Although the study was likely underpowered, benefits of the speech modality seemed to be outweighed by concerns of lack of control over the agent during a presentation, given the potential for the agent to barge in and take control if a participant paused for too long, and given concern over not remembering the exact phrases required to perform an explicit hand-off.

6 HAND-OFFS USING NONVERBAL CUES

Based on our observational studies, the most commonly-used hand-off cues were gaze and gesture at the next speaker. We felt that these would also be more acceptable cues since they comprise intentional selection of the next speaker by cue presence, rather than representing a default behavior the agent would take in the absence of a cue (i.e., the absence of speech during a long pause).

6.1 Implementation

Both the nonverbal hand-off cues of interest could be detected using the Microsoft Kinect sensor, with body orientation used as a proxy for user gaze direction. Based on our observational studies, we found that 'Sway Gesture' was the most commonly-used. We trained models using the Kinect Gesture Building tool to detect this and a second gesture to interrupt the agent.

The resulting system was then integrated into DynamicDuo. To improve accuracy and reduce false positives, a moving average technique was used to ensure the data from the Kinect was being reported accurately. Body orientation data was also integrated into the detection system, to ensure that the presenter was facing the co-presenter agent while gesturing to preclude false detections when gesturing at the presentation slides or towards the audience.

6.3 Evaluation

We conducted a two-session within-subjects study, identical in format to the one presented in Section 5, to evaluate the nonverbal hand-off detection system in comparison to the clicker in DynamicDuo. We also had judges review the recorded presentations to provide relative and absolute ratings of presentation quality.

We recruited 18 participants (4 male, 14 female, age 19-26, mean 22, SD 2.09) via an online advertisement. Of the 18 participants, 15 were categorized as low competence public speakers, and 3 as high competence public speakers based on the Self-Perceived Communication Competence Scale [23].

6.4 Presenter Study: Quantitative Results

All our results were normally distributed based on results of Shapiro-Wilk tests, with the exception of our assessment of Speaker Confidence scores. For analysis, we used repeated measure ANOVAs and t-tests for all of these measures with exception of Speaker Confidence, which was analyzed using Friedman rank sum tests.

6.4.1 State Anxiety

There was no significant difference between the two conditions for state anxiety (p=.181). However, the results suggest that participants were slightly more anxious prior to presenting with the gesture-based version (mean 31.5, SD 7.75) when compared to using the clicker version (mean 29.28, SD 7.27), possibly due to the novelty of the system.

6.4.2 Speaker Confidence

There was a significant difference in speaker confidence between the gesture and the remote control systems (χ2(1)=4.57, p=.03). Participants reported significantly higher confidence about their presentation when using gesture-based hand-offs compared to the remote control version.

6.4.3 Agent Ratings

There were no significant differences between agent ratings across the two conditions. Overall, ratings for both of the systems were highly positive (Table 2).

6.4.4 System Accuracy

System accuracy was assessed post-study by manually annotating the use of gesture commands in the recorded presentations, finding relatively high accuracy rates (mean 82.5%, SD 23.8%)

Table 2: Agent Ratings for Gesture vs. Clicker Interface

Ratings of Co-presenter: 1 - Not at all; 7 - Very Much	Gesture Mean (SD)	Remote Mean (SD)
How *satisfied* are you with…?	5.89 (1.49)	6.22 (1.17)
How much do you *like*…?	5.56 (1.62)	5.5 (1.38)
How much do you feel you	6.00 (1.28)	6.22 (1.00)
How *helpful* was…?	5.83 (1.54)	6.17 (0.99)
How much would you like to *give future presentations* with…?	5.56 (1.97)	5.67 (1.81)
How *easy* was it to use…?	4.44 (2.00)	4.12 (2.00)

6.5 Presenter Study: Qualitative Results

The analysis of semi structured interviews gave the following themes.

6.5.1 Audience Approval

Most participants thought the gesture-based interface would be more appealing to audience, e.g., *"It's more like a conversation interaction versus I am controlling the robot or this virtual thing"* (P5). They felt that the *"gestures would seem much better because it would be like having an actual co-presenter or a human co-presenter... It seems natural"* [P14]. Participants also indicated that the gestures would aid the audience in changing focus, because the motions would *"...move their attention towards the virtual assistant or co presenter, but while you are using buttons they won't even know that you're transferring the presentation to the co-presenter* " [P15].

6.5.2 Naturalness

Many participants felt that the gestural interface was a more conversational style of interaction, stating that it felt *"quite natural in that it is easier than just saying 'I am handing over to her'"* (P13)". Participants also reported that they felt *"like Angela is a real person because if it's a real person you would probably gesture to refer to other person"* [P20], and that they *"really liked the fact that I could turn and face her just like I would do if I was co-presenting with someone"* [P19].

6.5.3 Sense of Control

The biggest concern participants had around the gestural interface was the loss of precise control that came with the clicker. One participant stated that the remote version felt *"... like you can control it because if you gesture I'm kind of afraid that it's not going to work or I'll need to do it a second time or a third time*" [P20], and *"a button is so much more reliable"* [P10]. However, most participants still felt that *"the [gestural interface] was much better because it gave it a personalized touch"* [P15] and that *"it wasn't too much to remember. There weren't any complications as such, so it was pretty straight forward"* [P14].

6.6 Audience Perception Study

To evaluate the impact of these interfaces on audience's experience, we ran a follow-up study using videotaped presentations collected in the previous study of the gestural interface. Judges were asked to watch two pairs of videotaped presentations, with each pair comprised of the same speaker using either the remote control or the gestural interface. After watching, judges were then asked to compare the presentations based on: organization, note reliance, timing, pacing, and overall quality. Each item was judged using a 4-point ordinal scale of no difference, slight difference, moderate difference and substantial difference along with a field to indicate the better presentation. The ordering of presentations given to the judges was randomly assigned and counterbalanced. Twelve (3 male, 9 female) judges were recruited to compare the 12 pairs of recorded presentations.

6.6.1 Audience Perception: Quantitative Results

Wilcoxon signed-rank tests were used to evaluate the judges' ratings of the presenters and the presentation quality. Presenters using the gestural interface were rated significantly more exciting (p=.029) and entertaining to watch (p=.039), compared to those using the remote control. This result translated to overall presentation ratings, with a trend in overall presentation quality being higher for presentations with the gestural interface (p=.065). Upon further examination, we found that videos with participant needing to repeat the gesture more than once to hand-off (4 out of the 24 rated videos) severely degraded the judges' perception of presentation quality. When comparing ratings of presentations with 100% system accuracy to those without, the significance of the overall presentation quality comparison became statistically significant (p=.019), suggesting that accuracy is a key component.

6.6.2 Audience Perception: Qualitative Results

During post-study interviews, judges reported that presenters using the gestural interface seemed *"to be a lot more natural to me than just clicking a button. It reminded me of how when you are actually co-presenting, you gesture to the other person and so the audience knows what's going to happen*" [J1]. Some judges also felt that the gestural interface *"make the presentation more engaging, entertaining"* [J6].

7 Conclusions and Future Work

Results of our user studies provide support for use of natural speaker hand-off cues in human-agent presentations. While user acceptance of the speech-based hand-off system was generally poor, participants using the gestural interface were found to be more confident in their presentation, which in turn increased judges' perception of overall presentations. However, accuracy and reliability of the cue identification played a significant role in acceptance of these technologies, with multiple presenters voicing their concerns about the system failing during a presentation, and with judges stating that overall presentation quality significantly decreased when the system failed to work optimally.

DynamicDuo does not support improvisation since it requires preparation of a linear script in advance. However, once it can support dynamic presentations, the multi-modal turn taking system will become even more important since the agent can identify its parts of a presentation using both speech and gesture based cues. Also, the system does not currently provide a mechanism if recognition of the cues (speech or gesture) fails.

In our future work, we plan to explore the use of a wider range of verbal and nonverbal cues for speaker hand-offs. We also plan to further evaluate the reception of the system across a more diverse group of participants, since the majority of those participating in our study had relatively low self-rated presentation competence. Our ultimate goal is also to evaluate this technology in real public speaking venues.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under Grant Number IIS-1514490. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We thank Elise Masson and Teresa O'Leary for running our user studies.

REFERENCES

[1] Trinh, H., Ring, L. and Bickmore, T. DynamicDuo: Co-presenting with Virtual Agents. In *Proceedings of the CHI'15 Conference. (*2015).

[2] Bickmore, T., Trinh, H., Hoppmann, M. and Asadi, R. Virtual Agents in the Classroom: Experience Fielding a Co-Presenter Agent in University Courses. In *Proceedings of the Intelligent Virtual Agents*, 2016.

[3] Baudel, T., & Beaudouin-Lafon, M. (1993). Charade: remote control of objects using free-hand gestures. *Communications of the ACM*, *36*(7).

[4] Sommool, W., Battulga, B., Shih, T. K., & Hwang, W. Y. (2013, October). Using Kinect for holodeck classroom: A framework for presentation and assessment. In *International Conference on Web-Based Learning* (pp. 40-49). Springer, Berlin, Heidelberg.

[5] Noma., "A virtual human presenter." *IJCAI'97 Workshop on Animated Interface Agents*. 1997.

[6] Nijholt, "Introducing an Embodied Virtual Presenter Agent in a Virtual Meeting Room." *Artificial Intelligence and Applications*. 2005.

[7] Duncan, S. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3 (1974), 161-180.

[8] Goodwin, C. Achieving Mutual Orientation at Turn Beginning. Academic Press, City, 1981.

[9] Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26 (1967), 22-63.

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

[11] Duncan, S. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23 (1972), 283-292.

[12] Raux, A. and Eskenazi, M. Optimizing the turn- taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, 9, 1 (2012), 1-23.

[13] Ward, N., Fuentes, O. and Vega, A. Dialog prediction for a general model of turn-taking. In *Proceedings of the Interspeech* (2010).

[14] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, 41 (1998), 295-321.

[15] Gravano, A. and Hirschberg, J. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25, 3 (2011), 601-634.

[16] Hjalmarsson, A. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53, 1 (2011), 23-35.

[17] Thorisson, K. R. Gandalf. An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People. 1997.

[18] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. and Yan, H. Embodiment in Conversational Interfaces: Rea. *In: Proceedings of the CHI'99 Conference* (1999).

[19] Huang, L., Morency, L. and Gratch, J. Virtual Rapport 2.0. In *Proceedings of the IVA'11* (2011).

[20] Jonsdottir, G., Thorisson, K. and Nivel, E. Learning Smooth, Human-Like Turntaking in Realtime Dialogue. In *Proceedings of IVA'08* (2008).

[21] Chao, C. and Thomaz, A. Timed Petri nets for fluent turn-taking over multimodal interactoin resources in human-robot collaboration. *International Journal of Robotics Research*, 35, 11 (2016), 1330-1353.

[22] Jegou, M., Lefebvre, L. and Chevaillier, P. A. Continuous Model for the Management of Turn-Taking in User-Agent Spoken Interactions Based on the Variations of Prosodic Signals. In *Proceedings of the IVA'15* (2015)

[23] McCroskey, J. and McCroskey, L. Self-report as an Approach to Measuring Communication Competence. *Communication Research Reports*, 5 (1988), 108-113.

[24] Cassell, J., Vilhjálmsson, H. and Bickmore, T. BEAT: The Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH '01* (2001).

[25]Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I. (2006, May). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 1, pp. I-I).

[26] Spielberger, C. D. (1989). State-trait anxiety inventory: bibliography (2nd ed.). Consulting Psychologists Press.

[27] Paul G. L. (1966). Insight and desensitization in psychotherapy: An experiment in anxiety reduction. Stanford University Press.