Kevin Corti and Alex Gillespie

Offscreen and in the chair next to your: conversational agents speaking through actual human bodies

Article (Accepted version) (Refereed)

Original citation: Corti, Kevin and Gillespie, Alex (2015) *Offscreen and in the chair next to your: conversational agents speaking through actual human bodies.* Lecture Notes in Computer Science, 9238 . pp. 405-417. ISSN 0302-9743

DOI: 10.1007/978-3-319-21996-7_44

Offscreen and in the chair next to you: Conversational agents speaking through actual human bodies

Kevin Corti[1] and Alex Gillespie[1]

Abstract. This paper demonstrates how to interact with a conversational agent that speaks through an actual human body face-to-face and in person (i.e., offscreen). This is made possible by the cyranoid method: a technique involving a human person speech shadowing for a remote third-party (i.e., receiving their words via a covert audio-relay apparatus and repeating them aloud in real-time). When a person shadows for an artificial conversational agent source, we call the resulting hybrid an "echoborg." We report a study in which people encountered a conversational agent either through a human shadower face-to-face or via a text interface under conditions where they assumed their interlocutor to be an actual person. Our results show how the perception of a conversational agent is dramatically altered when the agent is voiced by an actual, tangible person. We discuss the potential implications this methodology has for the development of conversational agents.

Keywords: design criteria and design methodologies · evaluation methodologies and user studies · applications for film, animation, art and games · real-time interaction · cyranoid

1 Introduction

You have just signed up to participate in a social psychological research project and find yourself in a university meeting room with two chairs positioned facing each other. As you take a seat in one chair, the researcher informs you that the study you are partaking in involves holding a 10-minute face-to-face conversation with a stranger - another research participant. You are told that you can discuss anything you would like with your interlocutor; there is no script you need to follow or role you need to play. *Simple enough*, you think to yourself.

Shortly after the researcher leaves the room an ordinary-looking person enters and takes a seat in the chair across from you. You introduce yourself and reflexively utter "how are you?" to get the conversation started. "Fine," they say. As you begin to ask more questions you notice something slightly odd about the person you are speaking with. They are taking an unusually long time to answer you. *Should it* *really take someone several seconds to answer the question "are you a student?"* You brush this off as shyness and try to find a topic to discuss. "What do you think of the weather today?" you ask. "I think the weather is beautiful today," they respond (after several seconds of slightly awkward silence). *Really? This person's idea of beauty is icy rain and wind?* "But don't you hate it when the underground stations get all wet and slippery from the rainwater?" you counter. "Yes," they say. "What exactly do you find beautiful about it then?" you demand. Their answer: "I actually don't. Or at least prefer it when I wake up and don't remember much from them, which is actually almost on a regular basis." *Huh?* "Sorry?" you utter. Their reply: "I forgive you."

About halfway through the conversation you start to suspect that this is all an act (you are participating in a social psychology study, after all). *But they can't be giving scripted responses, I'm allowed to say and ask anything I want!* By the time the 10-minutes have ended you haven nearly given up all hope of achieving a meaningful interaction with the other person. You found it impossible to build upon discussion topics with them. As soon as you would begin to develop conversational sequences about, say, favorite books or movies, your interlocutor would change the topic or completely forget what they had said several turns prior. The most frustrating things in your mind were the misunderstandings and the inability, despite your best efforts, to resolve them.

The researcher returns to the room and your interlocutor exits. "So, how did it go?" the researcher asks…

2 Background

2.1 Motivation

For the last two years we have conducted basic social psychological research involving people interacting with conversational agent computer programs that speak through actual human bodies. Our goal has been the development of a new research tool: the "echoborg." An echoborg is a hybrid agent composed of the body of a real person and the "mind" (or, rather, the words) of a conversational agent; the words the echoborg speaks are determined by the conversational agent, transmitted to the person via a covert audio-relay apparatus, and articulated by the person through speech shadowing. Echoborgs can be used in everyday social scenarios as well as in laboratory environments rich in mundane realism. The purpose of exploring the possibility of such a tool stems from an interest in studying human-agent interaction under conditions wherein research participants are neither psychologically constrained nor influenced by machine interfaces. The echoborg can be thought of as a means of investigating the role of the tangible human body in altering how machine intelligence is perceived and interacted with.

2.2 Stanley Milgram's "cyranoid method"

The story of how we came to develop the echoborg concept begins with the unpublished work of Stanley Milgram. Milgram, who is well-known throughout psychology for his studies on obedience to authority [1], conducted a series of small experiments in the late 1970s that explored the possibility of creating a single interactive persona (a "cyranoid") from separate individuals [2]. He trained research confederates to speech shadow – an audio-vocal technique in which a person immediately repeats words they receive from a separate communication source [3]. Once trained, Milgram staged interactions wherein he and other research assistants conversed through these speech shadowers with research participants naïve to the fact that the person they encountered was being inconspicuously fed what to say by an unseen source. The shadowers contributed no words of their own to these interactions. Time after time, Milgram's participants failed to detect the manipulation, even in contexts involving extreme incongruity between source and shadower, such as when he sourced for 11- and 12-year-old children during interviews with panels of teachers. Milgram described these participants as having succumbed to the "cyranic illusion": a phenomenon he defined as failing to perceive when an interlocutor is not self-authoring the words they speak.

Speech shadowing is a straightforward technique that requires fairly little time to master. It can be accomplished by having the shadower wear an ear monitor that receives audio from either a recording or a live, spontaneously communicating source. The shadower listens to this audio and attempts to replicate the words and vocal sounds they hear as soon as they are perceived. Research has shown that shadowers can fluidly replicate audio stimuli at latencies as short as a few hundred milliseconds [4,5,6]. Shadowers instinctively mimic the gestural elements of their source, unconsciously adopting their source's accent, cadence, stress, emphasis, and so on [7,8]. Shadowing is not a cognitively demanding task. Trained speech shadowers, not having to *think* about what to say, can divert cognitive resources to other actions. For instance, while shadowing, one can focus on producing body language and an overall physical demeanor consistent with the words one finds oneself repeating.

2.3 Replicating Milgram

Milgram died in 1984 at the age of 51 having never formally published his cyranoid studies, and the method lay dormant within social psychology for over two decades. In recent years, however, the cyranoid paradigm has re-emerged in experiential art and interactive design research [9,10,11,12]. Inspired by this work, we set out to replicate Milgram's original pilots, which he outlined in a speech he prepared for an American Psychological Association conference in 1984 [2] and which are described in a biography authored by Blass [13]. Our interest was in vetting the utility of the method as a technique for investigating aspects of person perception and as a means of experiencing a transformed social identity.

We explored a basic cyranic illusion scenario in an initial study [14]. Participants in a control condition conversed for 10-minutes in unscripted scenarios one-on-one and face-to-face with an adult male research confederate. In a treatment condition, participants spoke with the same research confederate, who this time speech shadowed for a female source. Participants then filled out a questionnaire in which they were asked questions that gauged their suspicions regarding the communicative autonomy of the person they encountered. Participants were also thoroughly interviewed to gain a sense of their subjective impressions of the person they had encountered (the confederate). No differences between the conditions emerged. No participant in either condition believed that their interlocutor was being fed lines from a remote third-party, and very few participants held doubts as to whether their interlocutor was producing self-authored words. The fact that the conversations were unscripted (participants were told they could talk about whatever they wanted during the interactions) played a significant role in impressing upon the participants the feeling that the person they had encountered could not have been giving rehearsed responses.

Following our initial study, we decided to recreate Milgram's teacher panel interview scenario. We designed an experiment in which a 12-year-old boy and a university professor alternated sourcing and shadowing for one another in mock interview contexts [14]. Panels of three to five research participants were asked to interview a stranger for twenty minutes in order to gain a sense of their intellectual capacity. No scripts were used; participants generated their own questions and remarks. Following the interviews, participants were asked in a number of ways whether they had doubts as to the communicative autonomy of the person they had interviewed. The vast majority of participants believed that they had engaged with a person who was articulating their own self-authored responses. This provided us with evidence that the cyranic illusion was robust even in situations involving extreme incongruity between source and shadower.

3 Dreaming of electric sheep

Following our replication of Milgram's original pilots, and on the basis of what we observed in additional small-scale cyranoid studies [15], we decided to explore the possibility of a cyranoid composed not of two human beings, but one composed of a human shadower and a conversational agent computer program source (the most extreme source-shadower incongruity we could imagine). We fashioned the term "echoborg" to refer to this special type of cyranoid feeling that the term captured a person with mechanical elements (i.e., a cyborg) who conveys these elements through repeated speech (echoes).

The idea of the echoborg was largely inspired by the premise explored in Phillip K. Dick's famous novel *Do Androids Dream of Electric Sheep?* [16] (which was later adapted into the film *Blade Runner*). In the familiar story, a post-apocalyptic earth is partially populated by androids physically indistinguishable from actual human beings. One of the many thought experiments raised by the novel regards the role of belief in attributing an inner essence to an interlocutor, and the role perceiving a human body plays in implying a particular inner essence (namely, a human one).

3.1 Creating an echoborg

A standard cyranoid requires a means for the source to overhear the words being articulated by an "interactant" (Milgram's term for those who engage with a cyranoid, either naïvely or in full knowledge that they are doing so), as well as a means for the shadower to receive speech from the source in real-time. If it is the researcher's goal to construct a covert cyranoid, then the apparatus will have to be composed of devices that are not perceptible to the interactant. In our standard covert cyranoid apparatuses we use a contraption of interconnected radio transmission devices. From one room, the source speaks into a microphone that connects to an FM radio transmitter. The signal is transmitted to an adjacent room where it is picked up by a pocket radio worn by the shadower, attached to which is a neck-loop induction coil that is concealed by the shadower's clothing. The shadower wears a flesh-colored inner-ear monitor that sits in their ear canal and is not detectable at close distances. This monitor receives the signal from the induction coil, allowing the shadower to hear the source's words in real-time. A "bug" microphone placed in the room where the interactant and shadower are located wirelessly transmits audio via radio signal to a receiver listened to by the source, thereby enabling the source to hear and respond to the words of the interactant. While this amalgam of devices is convenient and inconspicuous, there are other means of constructing cyranoid apparatuses both overt and covert in nature [10].

The echoborg concept simply replaces the human source in a traditional cyranoid with a conversational agent of some sort (e.g., a chat bot). The means by which the agent receives speech from the interactant, and how it transmits its responses to the human shadower, are decisions that the researcher must make on the basis of their particular research objectives. One could opt for full technological dependency and make use of speech recognition software as a means of inputting the interactant's words into the conversational agent program, as well as speech synthesis software as a means of relaying the agent's words to the shadower. The advantage of full technological dependency is that it truly removes the human element from the echoborg's speech interpretation and speech projection subsystems. However, the downside of full technological dependency is that the quality of the interactions will be significantly constrained by current limitations in speech recognition and speech synthesis software. These technologies are not nearly as adept as humans at accurately perceiving spontaneous speech in real-time and articulating words with phonetic richness [17].

An alternative to full technological dependency is to have a human intermediary listen to the words spoken by the interactant (from a separate room), manually speed type them into the agent's input window, and speak the agent's subsequent response to the shadower via a radio transmission device (as the human source might in a standard cyranoid apparatus). The advantage of this minimal technological dependency format is that it preserves the verbal agency of the conversational agent (i.e., the agent still decides what to say in response to the interactant) while ensuring that the most accurate representation of the interactant's words gets interpreted by the agent.

4 Using echoborgs to study social perception: A simple comparative study

We ran an experiment to see how the experience of interacting with a conversational agent changes when the agent's words are embodied by a real person in face-to-face interaction. Three chat bots were used in the study, Cleverbot [18], Mitsuku (winner of the 2013 Loebner Prize chat bot competition) [19], and Rose (winner of the 2014 Loebner Prize) [20]. In the experiment, participants were *not* informed before the interactions commenced that they would be speaking with a conversational agent (i.e., the agents operated covertly). Participants either engaged a person who, unbeknownst to them, shadowed for a chat bot, or engaged who they assumed was another real person via a text interface. The study was approved by an ethical review board and was conducted in a behavioral research laboratory.

4.1 Participants

Forty-one adult participants (mean age = 24.12; 26 female) were recruited from a university recruitment portal and randomly assigned to one of two conditions (Echoborg or Text Interface) as well as a chat bot (Cleverbot, Mitsuku, or Rose). A female graduate student (aged 23) functioned as the echoborg's shadower. In the Echoborg condition, Cleverbot and Rose each spoke with seven different participants while Mitsuku spoke with six. In the Text Interface condition, the three chat bots each spoke with seven different participants.

4.2 Procedure

Echoborg condition. From the interaction room, the participant was informed that the study involved speaking to a stranger (another research participant) for 10-minutes and that they could discuss topics of their choosing during the interaction so long as nothing was vulgar. The researcher then left the room and relocated to an adjacent room which housed the computer on which the chat bot operated. The female shadower entered the interaction room and seated herself in a chair opposite the participant. The study made use of a minimal technological dependency format: as the participant spoke, the researcher speed typed their words into the chat bot's input window and articulated the chat bot's subsequent response into a microphone which discreetly relayed to the shadower's ear monitor (see Fig. 1). After 10-minutes, the researcher returned to the interaction room and the shadower exited.

Text Interface condition. From the interaction room, the researcher informed the participant that the study involved speaking to a stranger (another research participant) for 10-minutes. The participant was instructed that though they were being asked to speak aloud, the stranger's responses would appear on a computer monitor in the form of text. Participants were informed that they could discuss topics of their choosing. As with the Echoborg condition, the Text Interface condition involved a minimal technological dependency format: the participant's words were input by the researcher into the chat bot's input window; once the chat bot generated a response to the input text, the researcher routed the text response via an instant messaging client to the participant's screen (see Fig. 1).

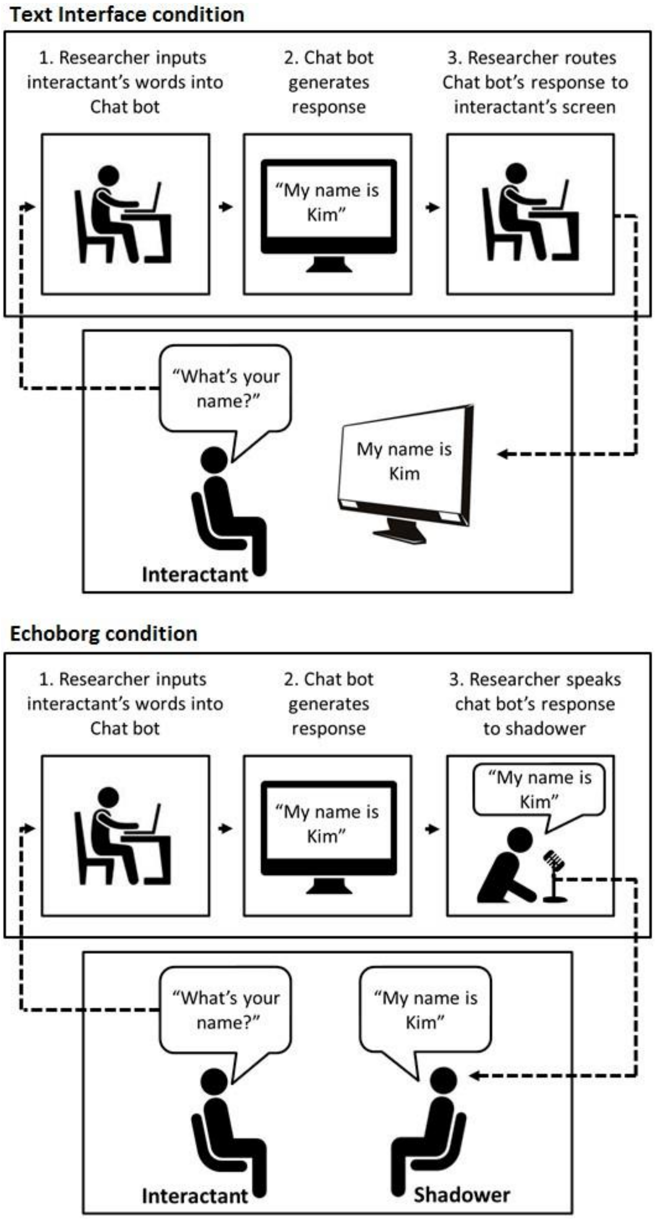

Fig. 1. Illustration of interaction scenarios used in study.

4.3 Measures and post-interaction interview

Following the interaction, the participant indicated on a 10-point scale how comfortable they felt during the interaction (1: not at all comfortable; 10: very comfortable) and also wrote a brief description of their interlocutor. Then, in a short interview, the participant was asked by the researcher to describe the personality of their interlocutor. Following the interview, the participant was debriefed and made aware of the full nature of the study. After all experimental trials were complete, participants' written descriptions were collated and adjectives and other descriptors used to describe the interlocutor's personality were identified. Adjectives and descriptors regarding personality were also extracted from the recorded post-interaction interviews.

4.4 Results

Bootstrapped independent samples means tests of participants' comfort ratings were conducted for each chat bot comparing those who engaged an echoborg to those who engaged a text interface. Participants who encountered Mitsuku via an echoborg felt significantly less comfortable than those who did so via text (mean difference = -2.57, *SE* = 1.02, 95% CI: [-4.58, -0.67]). Likewise, those who encountered Rose via an echoborg felt significantly less comfortable than those who did so via text (mean difference = -2.71, *SE* = 1.26, 95% CI: [-5.05, -0.04]). No significant difference between conditions was found among those who spoke with Cleverbot (see Fig. 2).

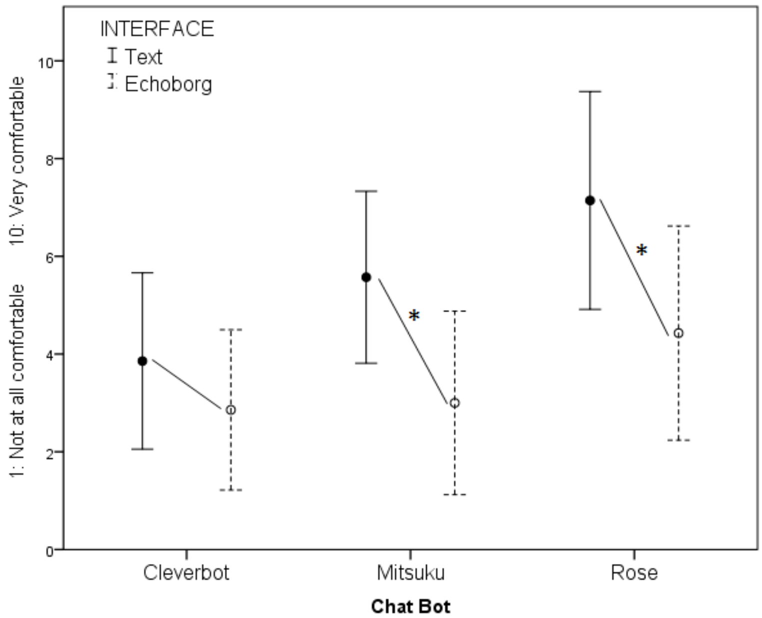

Fig. 2. Clustered error bar comparison of participants' feelings of comfort (*denotes significant difference: 95% CI of difference does not span zero).

Considering the written evaluations and post-interaction interviews, participants who spoke with an echoborg used a total of 86 unique descriptors to characterize their interlocutor, compared to 80 unique descriptors used by those who had spoken to a text interface. All descriptors which had a unique frequency of at least 3 (i.e., the descriptor was used by at least 3 participants in the same experimental condition) are shown in word clouds below (Fig. 3 and Fig. 4). The most frequent descriptors used in the Echoborg condition were “awkward” (6 different participants used this descriptor), “shy” (5), “introverted” (5), “uncomfortable” (4), “autistic” (3), “strange” (3), “poor social skills” (3), and “random” (3). The most frequent descriptors used by those in the Text Interface condition were “computer” (6), “strange” (5), “robotic” (5), “mechanical” (4), “introverted” (4), “difficult” (3), “friendly” (3), “random” (3), “nonsensical” (3), “odd” (3), and “asocial” (3).

Fig. 3. Personality descriptors used by participants in the Echoborg condition to describe their interlocutor. Word cloud shows descriptors that had a frequency greater than or equal to 3. Larger text size indicates greater relative frequency.

Fig. 4. Personality descriptors used by participants in the Text Interface condition

5 General discussion

5.1 Findings

Our results demonstrate that when words generated by a conversational agent become embodied by a real human body during face-to-face communication, the social psychological dynamics of the interaction dramatically alter. Keeping in mind that participants were not told until *after* the post-interaction interviews were complete that they had in fact been communicating with the words of a chat bot, participants who encountered a text interface were prone to use adjectives describing their interlocutor as artificial/inhuman (frequently using words such as "mechanical," "computer," and "robotic"). By comparison, those who encountered an echoborg more often used descriptors that pointed to intrinsically human characteristics (e.g., "shy," "awkward," and "autistic"). This is evidence of how the interface one encounters frames the interaction and functions as a prism through which meaning and perception are refracted. The tangible *face* of the Other evokes expectations regarding interpersonal dynamics and communicative norms that when violated trigger casual attributions tied to what is salient to a social actor. That is to say, when communication breaks down (e.g., due to an interlocutor lacking human-level discourse capacity on account of their words being generated by an artificial conversational agent), people look to what information is readily available to them in an effort to explain the breakdown; this is the principle of perceptual salience [21]. Participants who spoke with an echoborg based their personality judgements to what they saw: a human person sitting directly in front of them. Participants who spoke with a text interface based their personality judgments to what they saw: a computer screen.

We can also see how uncomfortable it is speaking to a covert echoborg relative to encountering a covert conversational agent via text. This again underscores the notion that face-to-face, in-the-flesh interactions place much higher intersubjective demand on the parties to an encounter. In future research we plan on further exploring the issue of interpersonal comfort in overt and covert echoborg contexts and linking our findings with the growing body of research surrounding "uncanny valley" phenomena emergent in human-agent and human-android interaction [22,23].

5.2 How Echoborgs can inform intelligent agent development

If a developer's goal is to create an embodied conversational agent that displays and elicits psychological responses identical to those that occur during human-human interaction, then an echoborg can be used to establish a benchmark against which the agent is evaluated. Comparative studies can be performed that observe users' reactions when the dialog component of an agent is projected through a human speech shadower and compare these observations to users' reactions when the same dialog component is projected through mechanical or virtual interfaces or avatars. Of course, in most instances it will be useful to inform research participants beforehand that their interlocutor is producing the words of an agent (i.e., an overt scenario as opposed to a covert scenario). This will attenuate the expectation of human-level discourse (as is the case when people speak with embodied conversational agents; people are under no illusions that they are speaking with a real human being when they do so). Covert scenarios, however, are interesting for the very reason that they set the bar quite a bit higher as they evaluate the dialog component of a conversational agent under conditions wherein participants expect an interlocutor fully capable of contextually appropriate human-level discourse.

Second, echoborgs are useful in basic social perception research akin to the study reported in the present paper. Personality judgments and other perception/attribution measures can be collected from interactants during and following interaction with a conversational agent via a variety of interfaces (e.g., text, onscreen avatar, mechanical android, offscreen human body, etc.) and under a variety of knowledge contexts (e.g., knowing vs. not-knowing their interlocutor is producing the words of an agent). This can give developers a sense of how perception of and interaction with a fixed intelligent agent change depending on the interface through which the agent communicates, from a minimal onscreen interface all the way up to a full-blown offscreen human body.

Third, as they have actual human bodies, echoborgs are mobile and capable of complex motor behavior. This allows for the possibility of non-stationary social interaction between research participants and conversational agents. At the moment, the agent's agency within an echoborg is limited to determining speech (the shadower still decides what motor behaviors to display), but this does not rule out the possibility of constructing agents that signal to their speech shadowers when to perform certain behaviors (e.g., using short tone patterns to indicate actions such as "shake hand," "stand up," "smile," and so on, or simply sending motor behavior instructions to their shadower's left ear monitor while sending what to say to their right ear monitor). Developing echoborgs of this nature would give researchers the ability to leapfrog current constraints on recognizing social actions (e.g., a gesture inviting a handshake) and motor coordination, allowing developers to focus on creating higher-level computer programs that can guide a human's behavior in social settings.

5.3 Conclusion

This paper has demonstrated how it is possible to construct scenarios wherein people encounter a conversational agent through the body of a real person offscreen. These agent-human hybrids are a special type of cyranoid we refer to as "echoborgs." We feel that, as a tool for interaction researchers and intelligent agent developers, the echoborg holds immense promise for furthering our understanding about how the human body shapes experiences with and perception of conversational agents.

References

1. Milgram, S.: Obedience to Authority: An Experimental View. Harper and Row, New York (1974)
2. Milgram, S.: Cyranoids. In: Blass, T. (ed.) The Individual in a Social World: Essays and Experiments, pp. 402-409. Pinter and Martin, London (2010)

3. Schwitzgebel, R., Taylor, R.W.: Impression Formation Under Conditions of Spontaneous and Shadowed Speech. Journal of Social Psychology 110, 253-263 (1980)
4. Marslen-Wilson, W.: Linguistic Structure and Speech Shadowing at Very Short Latencies. Nature 244, 522-523 (1973)
5. Marslen-Wilson, W.: Speech Shadowing and Speech Comprehension. Speech Communication 4, 55-73 (1985)
6. Bailly, G.: Close Shadowing Natural Versus Synthetic Speech. International Journal of Speech Technology 6, 11-19 (2003)
7. Pardo, J.S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E.: Phonetic Convergence in Shadowed Speech: The Relation Between Acoustic and Perceptual Measures. Journal of Memory and Language 69, 183-195 (2013)
8. Goldinger, S.D.: Echoes of Echoes? An Episodic Theory of Lexical Access. Psychological Review 105, 251-279 (1998)
9. Mitchell, R.: An In Your Face Interface: Revisiting Cyranoids as a Revealing Medium for Interpersonal Interaction. In: Wouters, I.H.C., Kimman, F.P.F., Tieben, R., Offermans, S.A.M., Nagtzaam, H.A.H. (eds.) Proceedings on the 5th Interaction Design Research Conference (SIDeR): Flirting With the Future, pp.56-59 (2009)
10. Mitchell, R., Gillespie, A., O'Neill, B.: Cyranic Contraptions: Using Personality Surrogates to Explore Ontologically and Socially Dynamic Contexts. In: Martens, J., Markopoulos, P. (eds.) Proceedings of the 2nd Conference on Creativity and Innovation in Design (DESIRE), pp. 199-210. ACM, New York (2011)
11. Pawlak, L.: The St. Unicorn's Trust [Performance Art]. Retrieved from http://www.mradamjames.com/#The-St-Unicorns-Trust (2009)
12. Raudaskoski, P., Mitchell, R.: The Situated Accomplishment (Aesthetics) of Being a Cyranoid. In: Melkas, H., Buur, J. (eds.) Proceedings of the Participatory Innovation Conference (PIN-C 2013), pp. 126-129. LUT Scientific and Expertise Publications, Lappeenranta (2013)
13. Blass, T.: The Man Who Shocked the World: The Life and Legacy of Stanley Milgram. Basic Books, New York (2004)
14. XXX: XXX. Journal of Social Psychology (2015)
15. XXX: XXX. Integrative Psychological and Behavioral Science (2015)
16. Dick, P.K.: Do Androids Dream of Electric Sheep? Doubleday, New York (1968)
17. Pieraccini, R.: The Voice in the Machine: Building Computers that Understand Speech. MIT Press, Cambridge (2012)
18. Carpenter, R.: Cleverbot [Computer Program]. Retrieved from http://www.cleverbot.com
19. Worswick, S.: Mitsuku [Computer Program]. Retrieved from http://www.square-bear.co.uk/mitsuku/housebot.htm
20. Wilcox, B.: Rose [Computer Program]. Retrieved from http://brilligunderstanding.com/rosedemo.html
21. Taylor, S.E., Fiske, S.T.: Point of View and Perceptions of Causality. Journal of Personality and Social Psychology 32, 439-445 (1975)
22. MacDorman, K.F., Ishiguro, H.: The Uncanny Advantage of Using Androids in Cognitive and Social Science Research. Interaction Studies 7, 297-337 (2006)
23. Seyama, J., Nagayama, R.S.: The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. Presence 16, 337-351 (2007)