SapientX Whitepaper

Avatars 101

Designing Avatars for Conversational AI.

By David Colleen

Introduction

When we began building conversational assistants (avatars), back in 2003[1], we knew intuitively that people were happier talking to characters rather than using a plain text or voice only interface. It would take years to reinforce our early instincts with hard science. This paper summarizes that science and what it means to the design of avatar based conversational systems.

Why Avatars

In our analysis, avatars make the driving experience better, safer and more satisfying. Dimitrios Rigas conducted a study comparing voice user interfaces with text and graphics vs. the same interface with an expressive avatar. He found a 4X improvement in users who rated the interface very good as well as a 3X reduction in those that rated the interface very poor. ("*The Role of Facial Expressions and Body gestures in Avatars for e-Commerce Interfaces*", Rigas & Gazepidis, 2009)

Many companies have begun to test avatars and have had positive results. "*photo-realistic avatars endowed with human-like voices in an online store induce increased feeling of trust, credibility, sociability and human warmth when compared with an online store with only images and text*." ("The Introduction of Avatars as a Factor of Sociability in e-Commerce Website", Alves & Soares,) "It has been known that visible verbal behavior (avatar animation) enhances the comprehension of verbal cues more, compared to voice only report." (Kang, et. al., 2016)

The Case For Avatars (the Eyes Have It)

In 1996, Blaxxun introduced a multi-user, avatar website called PointWorld (a precursor to Second Life). In this world, users could assume the body of an avatar and text chat with

1. The Sage project, part of the ARDA funded NIMD program.

other users worldwide. Users could choose and customize their avatar, selecting from humans, animals and even household objects. After several months of operation, founder Franz Buchenberger observed that visitors would only speak to other avatars if they had eyes.

This speaks to an innate human need to look at people when conversing. The reason for this is that much communication is passed with facial and body expressions (aka “back channel communication”) and that this non-verbal communication can only be received by looking at a person while talking to them. A voice assistant, without an avatar, does not fill this human need.


Blaxxun Pointworld

Socialization

Further, an avatar with eyes leads to more social user behavior. “*In the presence of a lifelike agent, people are more polite and tend to make socially desirable choices*” (Kramer et al., 2003)

In 2006, a study was performed using honor system payments. In the study, users were able to take a container of milk and voluntarily pay what they thought it was worth. When a picture of human eyes was displayed over the collection box, users voluntarily gave about twice as much money for the milk. The researchers concluded that a face, and specifically the eyes, gave the users a sense of being observed leading to higher levels of social responsibility. *We believe that images of eyes motivate cooperative behaviour because they induce a perception in participants of being watched.*” (Biol Lett, 2006 "Cues of being watched enhance cooperation in a real-world setting")

Lett went on to write, “*People tend to be generous, even toward unrelated individuals (Fehr & Fischbacher 2003). This is true even in situations where there is no prospect of repeat interaction, and hence no potential for direct reciprocity (Gintis et al. 2003). A possible mechanism maintaining generosity, where direct reciprocity is absent, is the motivation to maintain a pro-social reputation (Alexander 1987; Roberts 1998).*”.

Personification of Voice Assistants

My wife refers to SIri and Alexa as if they were humans. She treats them like not very intelligent children. Many of us do. I see this as a healthy and useful situation as users are also more forgiving of mistakes, when they see digital assistants as beings, which leads to higher user satisfaction.

A 2016 study of voice assistants, by the NHTSA, found that "*Users tend to blame themselves for non-optimal user/system interactions*" and that "*People seem to understand the limits of the technology. Users' expectations for system performance are modest – they tend to find some level of errors acceptable. User acceptance of voice assistants may be partially attributable to the tendency for users to blame themselves rather than the system for interaction errors.* "

Commands vs. Natural Conversation

Forcing users to learn commands can lead to frustration and unsafe driving conditions. "*Many in-vehicle (voice assistant) systems require specific voice commands. Such systems can lead to user errors when drivers forget commands, which is more likely with complex tasks.*" (NHTSA, 2016)

"*Human verbal communication patterns may serve as an effective model for interactions with voice systems. Users' in the study tended to anthropomorphize their voice systems. Therefore, it may be useful to consider making voice systems interactions more human-like in terms of turn taking behavior, use of natural speech patterns and vocabulary, and prosody. The design of efficient and highly acceptable interactions may benefit from a clearly defined human surrogate role and personality for the system (e.g., helpful personal assistant)*". (NHTSA study 2016)

The NHTSA study of drivers interacting with a variety of existing voice systems in cars found, "*One user said that he needed to speak loudly and clearly to the system, "almost as if you are talking to an idiot." Other participants said things like, "Sometimes she is stubborn," "She doesn't listen very well," and "What's wrong with you? Bad Navi!""*

It's important for a voice assistant to understand a user request in a natural conversation way. This leads to higher intent recognition, safety and user satisfaction. They should be able to address a voice assistant as if they were talking to a trusted friend. "*Users have a need for immediate and frequent system feedback in their interaction with VCS, as if in a conversation."* (NHTSA, 2016)

Existing voice systems, in cars, are purely command based. The longer a driver engages with a voice system, to perform a task, the more distracted they are consequently affecting their safety and user satisfaction. NHTSA reports the following duration times for voice control of common driving tasks:

	Avg. time w/errors	Avg. time no errors
Navigation destination entry	51 sec.	33
Communication	38	21
Information	28	22
Entertainment	23	14
Climate control	15	11

The average times, in the first column, are so long because the error rate of the existing voice systems are so high. JD Powers reports that existing automotive voice systems have a 63% error rate! These errors include failed speech recognition, failure to understand user intent and the user speaking the wrong command. The Society of Automotive Engineers (SAE) recommends a maximum of 15 seconds for safely conducted interactions.

We believe that these error rates can be significantly reduced and the interaction times lessened by using modern, high accuracy speech recognition (ASR) coupled with a fully conversational natural language system (NLU).

What we Talk About

In 2018, Brian Eberman, then CTO of Jibo, a home robot company, told me that 60% of their robots discussions with users were purely conversational. In his opinion, "*people desire companionship above all other functions*". Jibo's original programming was goal and activity oriented but Brian now seeks to add general conversational abilities.


Jibo home robot

Allen and Barbara Pease, in their 2001 book "*Why Men Don't Listen and Women Can't Read Maps*", document difference in male and female brain functions. They observed that, "*As adults, women talk about diet, personal relationships, marriage, children, lovers, personalities, clothes, the actions of others, work relationships and anything to do with people and personal issues. As men they discuss sports, their work, news, what they did or where they went, technology, cars and mechanical gadgets.*"

As designers of conversational systems, we need to allow for pure conversation, supported by topics that appeal to both men and women. So far, our assistants can talk on about 270 different topics including movies, astronomy and even jokes. Celeste Headlee, has written about what constitutes good conversations. I have included her top tips in Appendix A. We also think that it's important to include family entertainment (jokes, trivia and games) for long drives.

Variable Verbosity

We have found that some users like a talkative assistant where others want little or no verbal assistance. We address this by using word based sentiment analysis to continually monitor and score the users emotional state. Based on this, we can dynamically vary the

verbosity of our assistant. We can even hide or change the appearance of the avatar if the user dislikes it.

Still, it's important to convey that commands were understood and executed. Also, car makers want to deliver un-propted system warning and alerts to drivers. "*Confirmation steps were controversial. Some participants mentioned that confirmation steps were unnecessary or annoying, but several participants said that they liked having a confirmation step before placing a call so that they did not call the wrong number. One participant said that she didn't need to have a confirmation step because if she called the wrong number she could just cancel the call with a button press.*" (NHTSA, 2016)

Engaging Assistants

As we design digital assistants, we generally develop a "backstory" for the character that describes their style of interaction, their verbosity and if they pretend to be human or if they acknowledge that they are a computer program. We find that colorful characters are more fun and people more readily engage with them. This is supported by researchers such as Kang who found; "*People like virtual counselors that highly-disclose about themselves.*" (Kang and Gratch, 2007) We also advocate the user of celebrity likenesses and voices.

"*The results demonstrated that users reported more co-presence and social attraction to virtual humans who disclosed highly intimate information about themselves than when compared to other virtual humans who disclosed less intimate or no information about themselves. In addition, a further analysis of users' verbal self-disclosure showed that users revealed a medium level of personal information more often when interacting with virtual humans that highly-disclosed about themselves, than when interacting with virtual humans disclosing less intimate or no information about themselves.*" ("Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility", Cowell & Stanney, 2005)

What Should Avatars Look Like?

In the robot world, it has long been held that the best face for a robot was an abstract approximation of a human face. This was debunked in 2018 with the publishing of "*What People See in 157 Robot Faces*" by Evan Ackerman.
(https://spectrum.ieee.org/automaton/robotics/humanoids/what-people-see-in-157-robot-faces)

Ackerman found, in his study, a strong user preference for human faces over abstract robot faces. To us, this is not surprising since valuable communication information is carried in an expressive human face. Further, it was found that poorly crafted avatars, that displayed facial emotion, not in sync with the emotional state of the conversation, scored lower then better crafted avatars.

In 2018, we conducted our own study of user preferences for avatars. We found that 22% of respondents preferred a female avatar, 2% prefered male and 70% said they wanted a variety of avatars to choose from. In the study *"Cultures of Trust: Effects of Avatar Faces and Reputation Scores on German and Arab Players in an Online Trust Game*", researchers found that trust scores varied by 22% depending on the avatar appearance and that there was no differentiation based on gender. (Bente, et. al, 2014) This suggests that we need to do user testing to determine the most favored avatars.

Another research team wrote "*Thus, the attractiveness of an avatar should influence the likeability of an avatar, and likeability should mediate the degree of persuasion (ability).*" They found that a voice interface for shopping generated 15% higher user satisfaction and 21% higher entertainment value when an avatar was used. They also compared avatars that were visually optimized for attractiveness and compared them to avatars that were designed to look authoritative ("expert") and found the expert avatar was deemed 14% more expert in advice and 15% more credible but the attractive avatar was 4% more likeable. (Holzwarth, Janiszewski & Neumann, 2006)

Here are some of the avatars that we have developed for customers.

"*Our results indicate that while users generally prefer to interact with a youthful character matching their ethnicity, no significant preferences were indicated for character gender. For interaction, our results indicated that a character that portrayed trusting nonverbal behaviors was rated as being significantly more credible than a character portraying no nonverbal behavior, or one that portrayed non-trusting behaviors.*" (Cowell and Stanney, "Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents", 2003)

"*It was found that users prefer to interact with characters that match their ethnicity and are young looking. There was no significant preference for gender.*" (*The Application of Anthropomorphic Gamification within Transitional Healthcare: A conceptual framework,* Tuah and Willis, 2018)

Non-Verbal Communication

Albert Mehrabian, a pioneer researcher of body language in the 1950's, found that the total impact of a message is about 7 percent verbal (words only) and 38 percent vocal (including tone of voice, inflection, and other sounds) and 55 percent nonverbal.

Ray Birdwhistell pioneered the original study of nonverbal communication that he called "kinesics." Like Mehrabian, he found that the verbal component of conversation is less than 35 percent and that over 65 percent of communication is done nonverbally. Birdwhistell also estimated we can make and recognize around 250,000 facial expressions.

Gratch, Wang, et. al. at USC, ICT in their paper "*Creating Rapport with Virtual Agents*", 2007 made the case that AI-driven avatars, making use of body gestures (aka backchannel communication) were "*Overall, the current study and related findings add further evidence that the nonverbal behavior of virtual characters influence the behavior of the humans that interact with them. This gives confidence that embodied agents can facilitate social interaction between humans and computers, with a host of implications for application and social psychological research.*"

Charles Darwin (Universality of Facial Expression)

In 1872, Darwin published "*The Expression of the Emotions in Man and Animals*", in which he argued that all humans show emotion through similar behaviors. He wrote that emotion had an evolutionary history that could be traced across cultures and species. Today, many psychologists agree that many emotions are universal regardless of culture including: anger, fear, surprise, disgust, happiness and sadness.

We think that avatars, capable of understanding user emotions and that can respond emotionally, will generate a more successful user experience in conversational apps.


Charles Darwin

Emotion

Sensing the emotional state of a user and responding with an appropriate emotional state, both in voice and facial gesture, is important to successful and safe user interactions. “Subjects found it easier to attend to voice emotions similar to theirs. On average, drivers with the same emotion (as their voice assistant) had less than half as many accidents. “ (Nass & Brave, 2005)

“*Match and mismatch of emotion and voice tones, used by a voice assistant, with the driver’s current emotional state can have significant effects on driving behavior and distraction potential.*” (Nass & Brave, 2005).

Users that receive the proper emotional cues from an assistant are more satisfied and engaged. “*The results show that participants are sensitive to differences in the displays of emotion and cooperate significantly more with the cooperative agent.*” (Celso M. de Melo , Peter Carnevale and Jonathan Gratch "*The Impact of Emotion Displays in Embodied Agents on Emergence of Cooperation with People*", 2012)

Thomas and Johnston, 1995, "*emotional displays in an artificially generated character can have the general effect of making it seem human or lifelike, and thereby cue the user to respond to, and interact with, the character as if it were another person.*"

Facial Mirroring

We often subconsciously change our facial expression to mirror the expression of the person that we are talking to. Mirroring has been shown to help waitresses gain higher tips (Van Barren et al., 2003), enable sales clerks to achieve higher sales numbers (Jacob etc al, 2011) and give women more favorable ratings in speed dating (Gueguen, 2009).

Chris Frith, British neuroscience researcher, in his paper "*Role of facial expressions in social interactions*" wrote: “*The expressions we see in the faces of others engage a number of different cognitive processes. Emotional expressions elicit rapid responses, which often imitate the*

emotion in the observed face. These effects can even occur for faces presented in such a way that the observer is not aware of them. We are also very good at explicitly recognizing and describing the emotion being expressed. A recent study, contrasting human and humanoid robot facial expressions, suggests that people can recognize the expressions made by the robot explicitly, but may not show the automatic, implicit response. The emotional expressions presented by faces are not simply reflexive, but also have a communicative component. For example, empathic expressions of pain are not simply a reflexive response to the sight of pain in another, since they are exaggerated when the empathizer knows he or she is being observed. It seems that we want people to know that we are empathic. Of especial importance among facial expressions are ostensive gestures such as the eyebrow flash, which indicate the intention to communicate. These gestures indicate, first, that the sender is to be trusted and, second, that any following signals are of importance to the receiver."

We plan to add video input to our system to do mirroring and sentiment analysis. It will also allow us to capture micro expressions in users' faces.

Micro Expressions

What are micro expressions? Micro expressions are facial expressions that occur within 1/25th of a second. They are involuntary and expose a person's true emotions. They can happen as a result of conscious suppression or unconscious repression. These facial expressions are universal, meaning they occur on everyone around the world.

Avatar Gaze

Research shows that an avatar, that occasionally looks away, increases user engagement and gives a context to social interaction (Garau 2003, "*The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment*"). Further, an avatar that glances to the side, to acknowledge a new graphic

being displayed in the user interface, automatically leads the viewer to glance at the new graphic.

British neuroscience researcher, Chris Frith, in his paper "*Role of facial expressions in social interactions*" observed, “This result suggests that our tendency to follow the gaze direction of others is automatic and difficult to suppress.” (Kulms , Krämer, et. al. "*It’s in Their Eyes: A Study on Female and Male Virtual Humans’ Gaze*", 2011)

Frith wrote; “*In a 2x2 between subjects experiment we manipulated the Rapport Agent’s gaze behavior and its gender in order to test whether especially female agents are evaluated more negatively when they do not show gender specific immediacy behavior and avoid gazing at the interaction partner. Instead of this interaction effect we found two main effects: gaze avoidance was evaluated negatively and female agents were rated more positively than male agents.*”

We address this in our own avatars by making them blink, occasionally turn away from the viewer and look towards graphics that are displayed on the screen. With a microphone array and beamforming software, we can even make the avatar face the user that they are speaking to.

Trust

Voice systems that use an avatar generate more trust in users. We would suggest that trust is an important part of user satisfaction and critical in ADAS situations. “*They can exhibit greater trust of the agent’s recommendations*” (Cowell and Stanney, Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents, 2003); “*and they can feel more empathy*”. (Paiva et al., 2004). “*People are more comfortable revealing sensitive information to computers than face-to-face interviewers*” (S. Weisband and S. Kiesler, 1996. “Self-disclosure on computer forms: Meta analysis and implications”) " *the results from this study suggest that there may indeed be a benefit to endowing computer characters with nonverbal trusting behaviors, as long as those behaviors are accurately and appropriately portrayed*."

Recently it was also shown that virtual humans can increase willingness to disclose by providing a “safe” environment where participants don’t feel judged by another human interlocutor (Lucas et al. 2014). “*The preliminary findings indicated that interactants revealed greater intimate information about themselves in interactions with virtual humans than with*

real humans." (Kang and Gratch, "*The effect of avatar realism of virtual humans on self-disclosure in anonymous social interactions*", 2010) and "*Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories.*" (Kang SH, Gratch J., Studies in Health Technology and Informatics.)

"*Researchers have found that individuals who have low propensity to trust characterize a computer character as less credible compared with those with higher propensity to trust*", (On Manipulating Nonverbal Interaction Style to Increase Anthropomorphic Computer Character Credibility, 2002)

Driver Distraction

We have yet to see a published study that measures potential driver distraction when using an avatar in a car head unit display. In our view, an animated avatar has far less active animation than the moving map, navigation systems that are already in our cars. That said, we want to couch our best practices in industry standards and recommendations and work towards a safer and more rewarding driving experience.

In 2016, the US DOT, National Highway Traffic Safety Administration (NHTSA) published a guideline paper titled "*In-Vehicle Voice Control Interface Performance Evaluation*". In that paper, there were several relevant key points:

- Voice control systems may enable drivers to keep their hands on the steering wheel and their eyes on the road, and therefore may be less distracting than systems that demand visual-manual interaction.
- Drivers are more likely to look away from the road when engaged in voice-based tasks (and) voice interactions might produce cognitive load that is not reflected in glance behavior.
- In the case of a (voice-based) navigation destination entry on a production vehicle, the supposedly hands-free and eyes-free operation led to an average task completion time of over 90 seconds and an average of over 30 seconds of off-road glance time.

- The SAE Recommended Practice J2364 (SAE, 2004) standard recommends a 15-second maximum task completion duration, specific to visual-manual interfaces to minimize excess task completion times that could result in greater numbers of off-road glances. The standard was not intended to prescribe a safe limit of interaction.
- Speech recognition performance is critical because failures to understand drivers' commands increase the distraction potential of the system relative to error-free performance.
- Unfortunately most in-vehicle voice systems require the user to push a button before using a voice system. These push-to-talk systems help reduce recognition errors, but can increase drivers' workload (Fodor et al., 2012).
- A study by Reimer and Mehler (2013) found that the use of a voice assistant for address entry navigation task resulted in only 13.3 percent of participants conforming to the total eyes-off-road duration (less than 12 seconds) criterion of the visual-manual guidelines.

Our takeaway is that a proper NLU based voice system, combined with a well designed avatar, will improve drivers safety.

Conclusions

The first generation of voice interfaces were used for simple IVR tasks like selecting options in a voicemail system. Siri and Alexa were second generation systems that expanded available vocabulary selections but that still relied on keyword commands. We are in the midst of the third generation of voice systems that strives for broad domain, natural user input. Many of the first and second generation voice platforms have attempted to extend their systems to achieve third generation goals... with limited success. In our view, true natural language understanding can only be achieved by systems built, from the ground up, to do so.

British psychologist Elizabeth Stokoe wrote; "*In our view, consumers want and need digital interfaces that can readily understand what they say and respond with appropriate, meaningful voice response.*"

This paper looks beyond, to fourth generation voice systems capable of engaging users both with audio and visual communication. For cars, we believe this means a safer and more satisfying driving experience.

Last year, in a meeting with GM Cruise, we were stopped in the middle of our presentation and asked why we were talking about automotive controls. They went on to say that they are building an automated taxi system that has more in common with your living room than your car and that a person getting into such a taxi could not be expected to know about physical controls. In their view, a voice interface was the only logical choice!

Future Work

- Appraisal Theory
- Prosody
- Facial Recognition

Appendix A - Conversational Guidelines

To design a voice assistant to be good at conversation, we need to study what makes good conversation as well as how social “spark plugs” initiate discussion with strangers and make people that they are having an engaging conversation. Celeste Headlee, in her Ted Talk, offered the following guidelines to being a good conversationalist:

1. *Don't multitask. And I don't mean just set down your cell phone or your tablet or your car keys or whatever is in your hand. I mean, be present. Be in that moment. Don't think about your argument you had with your boss. Don't think about what you're going to have for dinner. If you want to get out of the conversation, get out of the conversation, but don't be half in it and half out of it.*
2. *Don't pontificate. If you want to state your opinion without any opportunity for response or argument or pushback or growth, write a blog.*
3. *Use open-ended questions. In this case, take a cue from journalists. Start your questions with who, what, when, where, why or how. If you put in a complicated question, you're going to get a simple answer out. If I ask you, "Were you terrified?" you're going to respond to the most powerful word in that sentence, which is "terrified," and the answer is "Yes, I was" or "No, I wasn't." "Were you angry?" "Yes, I was very angry." Let them describe it. They're the ones that know. Try asking them things like, "What was that like?" "How did that feel?" Because then they might have to stop for a moment and think about it, and you're going to get a much more interesting response.*
4. *Go with the flow. That means thoughts will come into your mind and you need to let them go out of your mind. We've heard interviews often in which a guest is talking for several minutes and then the host comes back in and asks a question which seems like it comes out of nowhere, or it's already been answered. That means the host probably stopped listening two minutes ago because he thought of this really clever question, and he was just bound and determined to say that. And we do the exact same thing. We're*

sitting there having a conversation with someone, and then we remember that time that we met Hugh Jackman in a coffee shop.

5. *If you don't know, say that you don't know. Now, people on the radio, especially on NPR, are much more aware that they're going on the record, and so they're more careful about what they claim to be an expert in and what they claim to know for sure. Do that. Err on the side of caution. Talk should not be cheap.*
6. *Don't equate your experience with theirs. If they're talking about having lost a family member, don't start talking about the time you lost a family member. If they're talking about the trouble they're having at work, don't tell them about how much you hate your job. It's not the same. It is never the same. All experiences are individual. And, more importantly, it is not about you. You don't need to take that moment to prove how amazing you are or how much you've suffered. Somebody asked Stephen Hawking once what his IQ was, and he said, "I have no idea. People who brag about their IQs are losers."*
7. *Try not to repeat yourself. It's condescending, and it's really boring, and we tend to do it a lot. Especially in work conversations or in conversations with our kids, we have a point to make, so we just keep rephrasing it over and over. Don't do that.*
8. *Stay out of the weeds. Frankly, people don't care about the years, the names, the dates, all those details that you're struggling to come up with in your mind. They don't care. What they care about is you. They care about what you're like, what you have in common. So forget the details. Leave them out.*
9. *This is not the last one, but it is the most important one. Listen. I cannot tell you how many really important people have said that listening is perhaps the most, the number one most important skill that you could develop. Buddha said, and I'm paraphrasing, "If your mouth is open, you're not learning." And Calvin Coolidge said, "No man ever listened his way out of a job." Why do we not listen to each other? Number one, we'd rather talk. When I'm talking, I'm in control. I don't have to hear anything I'm not interested in. I'm the center of attention. I can bolster my own identity. But there's another reason: We get distracted. The average person talks at about 225 word per minute, but we can listen at*

up to 500 words per minute. So our minds are filling in those other 275 words. And look, I know, it takes effort and energy to actually pay attention to someone, but if you can't do that, you're not in a conversation. You're just two people shouting out barely related sentences in the same place.

10. *Be brief. A good conversation is like a miniskirt; short enough to retain interest, but long enough to cover the subject.*