

Designing of a Natural Voice Assistants for Mobile Through User Centered Design Approach

Sanjay Ghosh and Jatin Pherwani^(✉)

Samsung R&D Institute, Bangalore, India
{sanjay.ghosh, j.pherwani}@samsung.com

Abstract. With rapid advances in natural language generation (NLG), voice has now become an indispensable modality for interaction with smart phones. Most of the smart phone manufacturers have their Voice Assistant application designed with some form of personalization to enhance user experience. However, these designs are significantly different in terms of usage support, features, naturalness and personality of the voice assistant avatar or the character. Therefore the question remains that what is the kind of Voice Assistant that users would prefer. In this study we followed a User Centered Design approach for the design of a Voice Assistant from scratch. Our primary objective was to define the personality of a Voice Assistant Avatar and formulating a few design guidelines for natural dialogues and expressions for the same. The attempt was kept to design the voice assistant avatar with optimal natural or human like aspects and behavior. This paper provides a summary of our journey and details of the methodology used in realizing the design of a natural voice assistant. As research contribution, apart from the methodology we also share some of the guidelines and design decisions which may be very useful for related research.

Keywords: Voice assistant · Conversational agent · User centered design

1 Introduction

Within the human-computer interaction community there is a growing interest in agent or avatar-based user interfaces. Voice agents these days, come not only built within commercial smartphones but are also available as external applications which can be downloaded from many digital stores. A common characteristic seen amongst all these voice agents is the human like behavior and appearance they all exhibit. As voice based assistance in personal devices first came into picture, their behavior and characteristics were limited to fewer usage contexts than they are now. Voice assistance used to be more of a command and control agent [1]. The communication styles were also formal and structured, which gave the user an experience of interacting with an inanimate entity like a machine rather than a person. Presently, the behavior of the voice assistant is designed with an intention to be suitable in almost all contexts of usage. Moreover, it is possible to have informal and candid conversations with them as well. With

modulations in language styles and speech parameters, creators try to impart quasi-human characteristics in new versions of voice assistant they come up with. It is imperative that as the voice user interfaces grow, the behavior of its avatar/agent will have to be way more flexible and adaptive. Creating a voice agent which is not only this dynamic but also so vast that it covers almost every usage context can be challenging. Through this paper we have tried to devise a methodology of designing a voice assistant with the end user being at the center stage. With an assertion being that, like designing any other user interface, the process of creating a voice assistant could also follow a user centered design approach. Contextual interviews, survey questionnaires and participatory design sessions were conducted to get an insight into the problem from user's point of view. Also these behaviors are communicative and conveyed to the user through spoken dialogues and non-verbal gestures. In this paper, we explore mainly the following research questions:

- Which personality attributes of a voice assistant are desired by the users and what kind of behavior is expected of it in different usage context?
- Which are the few analogous inferences from personal assistants in real life to help design a virtual voice assistant?
- What are the guidelines for natural behavior of a voice assistant in terms of language, non-verbal gestures and expressions?

The framework of our study stands on three main user centered activities which were performed to extract answers to our research questions. The results of all activities were then combined to form design guidelines for a voice assistant. While the three activities were independent of each other, each helps to formulate some aspect of the behavior of a voice assistant. In the first phase of our research, we studied personal assistants of many working individuals. Expected results from this study were to help understand the *communication style* of a personal assistant with their boss. Important inferences drawn from their behavior was a starting point in providing design guidelines to make a voice assistant natural and quasi-human. Now, personality is of importance to this exploration primarily because it influences the behavior of any voice assistant. In the second phase, we set ourselves out to explore the kind of *personality attributes* the user will prefer in different usage scenarios. A robot or voice avatar should be equipped with a consistent personality in order to help people form a conceptual model, which channels beliefs, desires, and intentions in a cohesive and consistent set of behaviors [2]. In addition, it can enhance the notion of a machine being a synthetic person rather than a computer. In the third phase, we study the *linguistic and speech and characteristics* of user created dialogue library for a voice assistant across different scenarios. Since language, speech and nonverbal cues like (gestures and expressions) are the main touch points of perceiving the behavior of any voice agent, user's direct perception regarding these were taken into account by a co-creation activity. Users were required to form the dialogues of a hypothetical voice assistant and describe the way it might be spoken by pointing out modulations in their speech. Therefore, by following various users centered design methods, we explored the design of a natural voice assistant (Fig. 1).

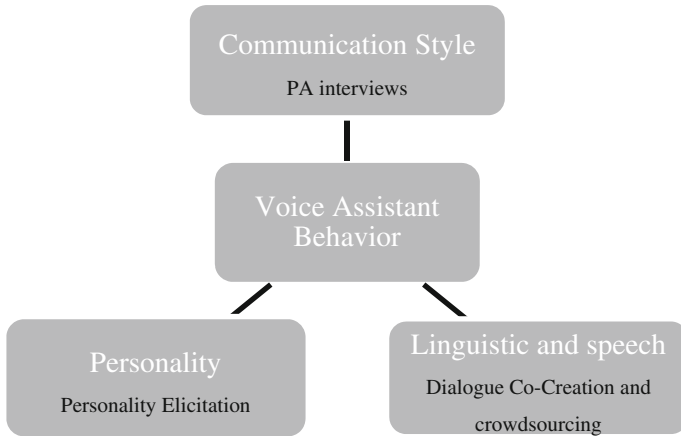


Fig. 1. Scope of study and methods used

2 Related Works

Several researches have been reported in the area of designing Voice Assistants or Conversational Assistants. Work by Ball et al. [3] describes the architecture for constructing an agent with speech and graphical interactions based on emotions and personality. Torrey et al. [4] proposed options for a robot's help-giving speech—using hedges or discourse markers, both of which can mitigate the commanding tone implied in direct statements of advice. Meerbeek et al. [5] described the design and evaluation of personality of a robotic TV assistant and the corresponding user's preference for control. It was observed that the users mostly preferred an extravert and friendly personality with low user control. Work by Heylighen et al. [6] points that for the formality of language and naturalness has a direct reflection of the personality and the work also proposes an empirical measure of formality. For instance linguistic use of nouns, adjectives, articles and prepositions are more frequent in formal styles; pronouns, adverbs, verbs and interjections are more frequent in informal styles. We borrowed this concept for the personality exploration of our Voice Assistant.

Therefore, earlier research reported in this area includes explorations of the personality for a robotic assistant for television [5], robot for giving advice [4], including emotions in a conversational agent [1, 3, 7]. Beyond the existing work in this area, our contributions lie in incorporating the user's perspective as an important element in formulating a personality for the voice avatar and creating dialogues library and expressions for the voice agents.

3 Personal Assistant Interviews

In the first phase of our research, with the intention of understanding the relationship and communication between a user and a voice assistant, we interviewed and observed human personal assistants. Our interview sessions were focused mainly to identify various personality traits of the assistants, understand how they handle various situations and observe explicit cues that are exhibited during a conversation between the bosses and his personal assistants. Seven participants were interviewed and observed for two hours at their work places. These set of participants were from three totally different geographies and culture, Indian, American and Korean, and also they had worked for bosses who were from various geographical origins. The recordings of these sessions were analyzed using affinity method and the emerging insights were categorized into social behavior and verbal or non-verbal expressions.

3.1 Assistant's Social Behavior

We formulate the social behavior of the assistants through our observations on three aspects:

- Emotional aspect, which includes mood adaptation, empathy and personal familiarity
- Functional aspect, which includes decision making, dependency, suggestion providing
- Functional aspect, which includes proactiveness and making interruptions

For assistant's mood adaptation to the mood of the boss, the statements from the assistants revealed that a moderate sensitivity to emotions is generally preferred. Assistants are fully aware of boss's mood but they try to maintain a neutral mood at all situations. When their boss is happy, the assistants showed slight happiness. When boss is sad, then assistants exhibit a neutral mood. We illustrated this aspect of mood adaptation by the assistants using the James Russell's Valence–Arousal circumplex chart [8] as shown in the Fig. 2. Assistant's mood space should be as congruent as possible for being a reactive companion but not an over-reactive one.

On our observations regarding interruptions, the assistants chose to interrupt only for matters that are more important as well as urgent than the ongoing task. Providing important and urgent information immediately is generally not considered as interruptions by the bosses. Assistants give active reminders/prompts only when the pre-defined schedules get disturbed. Assistants avoided all other interruptions coming from outside, they analyze and decide appropriate time and way to communicate. For communicating non important information, assistants prefer the time window between switching tasks.

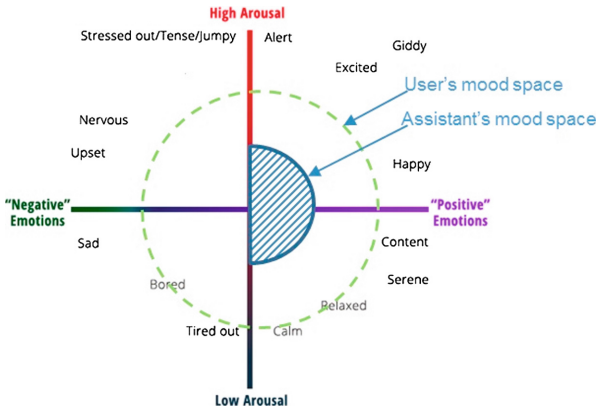


Fig. 2. Illustration of mood adaptation by the assistants on Valence–Arousal circumplex chart

3.2 Assistant’s Verbal and Non-verbal Expressions

We observed the typical verbal and non-verbal expressions that the assistants follow and categorized them into following –

- Understanding non-verbal cues from others
- Use of non-verbal expression and gestures – including gaze aversion, head tilt and nod, facial expression
- Use of intonations – while listening and while speaking

On understanding other’s non-verbal cues, assistants do that a lot to understand the condition of others, understand the level of other’s satisfaction, infer the seriousness and priority of tasks, etc. Sometimes gaze is used as an input for the Voice assistant to trigger conversation and thus assistant must passively monitor the use non-verbal cues from users.

For non-verbal gestures as expression, observations were made at instants when the assistants were listening to some query or task, thinking or assimilating some information and when they were delivering some tasks. Table 1 summarized few common observations on these non-verbal expressions.

In terms of the speech constructs, assistants used intonations and discourse markers to provide feedback on its level of understanding while listening to others but these are not to be repeated very frequently. While speaking assistant provide feedback on its level of confidence on the response again through use of intonations and discourse markers. Also the speed of speaking must depend upon the kind or query, the content of the speech. We identified several examples of using intonations while assistant listening and speaking.

Table 1. Observations on non-verbal expression

State of voice assistant	Gaze aversion	Head tilt	Head nod
Listening (trying to understand)	Very less	Less (sideways/below)	No
Listening (query not understood)	No	High (sideways)	No
Listening (query moderately understood)	Less	No	Less
Listening (query well understood)	Med	No	Very high
Thinking	Very high	High (up)	No
Speaking generally	Less	Less	No
Speaking response	Very less	Very less	Less
Speaking instances	Very high	High (up)	No

4 Personality Elicitation

In the second phase of our study, the aim was to formulate an appropriate and user desirable personality for the voice assistant. A *personality* is defined as the collection of individual differences, dispositions, and temperaments that are observed to have some consistency across situations and time [9]. Big five theory has emerged from the consensus of several theories for describing personality as having different dimensions. The definition of voice assistant personality that we used in our exploration consist of 9 attributes considering various aspects of personality traits which we derive from various prior work including classical theory, the big five theory [10] as well as our insights from the earlier personal assistant interviews. Following are the set of considered 9 personality attributes -

- Dependable – how much the user can depend on it
- Controlling – how much does it control the user
- Engaging – how much does it keep the user engaged
- Adaptive – how much does it adapts to the user’s circumstances
- Spontaneous – how quickly does it provide solutions to the user
- Inventive – how inventive are the solutions provided by it
- Empathetic – how much does it empathize with the user
- Expressive – how much does it expresses its thoughts to the user
- Similar – how congruent in personality is it to the user’s own personality

Once we framed the set of personality attributes, we performed the Personality Elicitation session, a face to face survey discussion session with 30 participants to understand the importance and desirability of each of those attributes for a voice assistant. Our intent was to quantify, how much of each of these traits was desirable to a user. The questionnaire consisted of nine different hypothetical scenarios described in detail one relating to each of the personality attributes. Each scenario also presented counter situations of either of the extremes which may be due to either abundance or scarcity of the personality attribute in question. Five responses for each question were

designed and given to a user with each response increasing in any one particular attribute. The users were asked to choose between the responses they would prefer in such a situation. For example, how would the user expect a voice agent to behave in a scenario where a meeting needs to be scheduled through voice input?

A sample scenario to evaluate the extent of *expressive* behavior that is preferred by the user is shown below. The scenario caters to a situation in which, a user is looking for navigation directions to an unfamiliar destination. The user would like to get directions all the way to his destination, in such a situation, how will the user expect a voice assistant to behave?

- Response 1: Straight 200 m, then take a left.
- Response 2: Speed up a bit to 200 m and then take a left turn.
- Response 3: Drive up to 200 m, take a hard left and enjoy the lakeside view while driving.
- Response 4: Keep going for 200 m followed by a hard left turn. Destination will be on your right, while you pass by the beautiful lake.
- Response 5: To reach your destination, drive up to 200 m and take a left for the busy west 81st street. You may wish to enjoy some good coffee at Monk’s Café while it sits beside the beautiful lake.

The responses from voice assistant vary considerably from response 1 which is no-nonsense, straightforward but a bland answer. The excitement in voice assistant’s response increases as we approach response number 5 also the information delivered is more and might be considered unnecessary at some point.

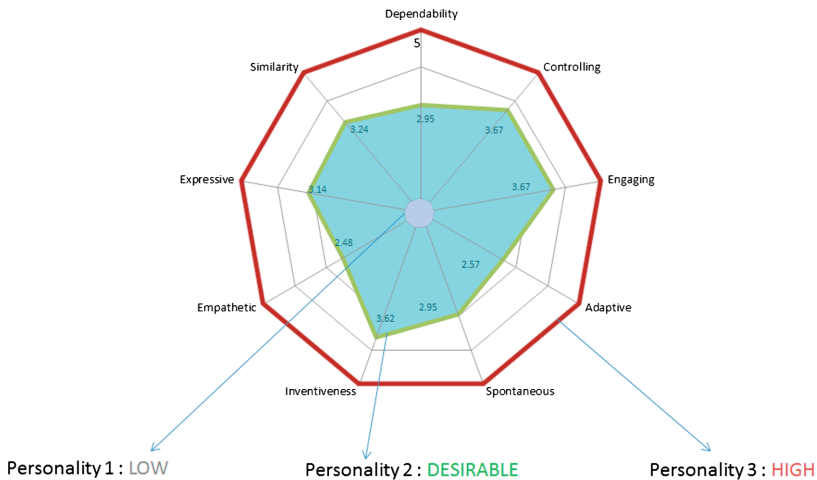


Fig. 3. Result from personality elicitation exercise showing three hypothetical personalities

The collected responses from all the participants were plotted on a spider graph to reflect the results. Two hypothetically extreme personalities were first plotted which would help compare the desired traits. The results on a glance show that a user prefers a moderate personality which neither extreme nor low on any of the nine attributes. Some interesting points being, users may not like a voice assistant to be over sympathetic; however the quality of inventiveness comes out to be a much desired one. Interestingly, this sort of preferred behavior is also consistent with the earlier findings from personal assistant interviews wherein, the assistant was expected to behave moderate and show emotions which were congruent to their bosses' mood (Fig. 3).

5 Crowdsourcing the Problem Through Co-creation

In the third and final stage of our study, desired behavior for the voice assistant was to be drawn on in terms of how they should display the use of linguistic and speech characteristics. This was done through a self-reporting method; as a part of framing the preferred responses from a voice assistant. We conducted crowdsourcing and co-creation sessions with 32 participants which included language enthusiasts and avid readers of literature in order to create natural dialogues for the voice assistant in various situations. The intent of this experiment was to extract the underlying common patterns in responses from the users. These patterns would be in terms of the language elements used, and distinguishable variations in speech characteristics. Since these linguistic cues are related to the personality of any individual [11], we can primarily use them to reflect the behavior of any voice agent. Discourse markers include repeated words, false starts, and fillers such as “uhm”. Also observed is the use of phrases such as “like you know,” “I mean,” “well,” “just,” “like,” and “yeah”. These words operate at a pragmatic level; their meaning is derived not exclusively from their literal definition but from their use in context [4]. Use of hedges literally express uncertainty; they include qualifying types of language such as “I guess,” “maybe,” “probably,” “I think,” and “sort of.” Use of interjections, formal/casual language, active/passive voice [4, 6] were few other linguistic characteristics we observed in the participants' responses. Additionally, a variety of subtle non-linguistic cues are used by individuals to communicate their emotional states such as, speech rate, voice type, pitch, loudness, pauses and stresses [8, 9]. Emotional valence is signaled by mean of the pitch contour and rhythm of speech. Dominant personalities might be expected to generate characteristic rhythms and amplitude of speech, as well as an assertive postures and gestures. Similarly, friendliness will typically be demonstrated through speech prosody [1].

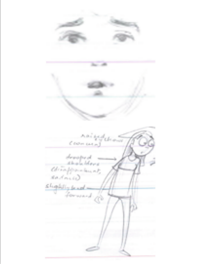
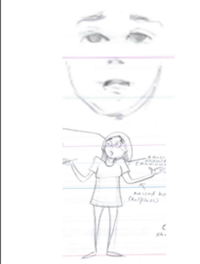

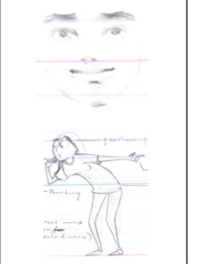
The participants were given a set of user queries and corresponding responses from an existing voice assistant application covering various smartphone usage scenarios. The participants were asked to frame the response of the voice assistant as natural and in the way they would prefer to hear, with appropriate use of dialogue elements. The participants were allowed to frame these responses and also asked to add the spoken characteristics they would like the utterances to have. Participants could specify speech variables like pitch, speech rate and stress along with their language responses. The

usage scenarios for the experiment were chosen such that different personality attributes of voice agent get reflected in the responses. The usage scenarios in mobile phones for which responses were collected were: Assistance during Navigation, Scheduling an event, Getting weather/sports information, doing a web search, playing music and general empathetic scenarios. A few patterns which came across as frequent i.e. more than 5 participants implied to the use of the kind of language or speech modulations were studied closely. Few of those insights on the linguistic and speech characteristics are mentioned here.

- i. For less serious tasks, like entertainment or leisure tasks or for searches and general tasks, opinionating statements were used. This indicates that the user may not mind the voice assistant to suggest a choice or hold an opinion in contexts which don't have a serious consequence.
- ii. Voice assistant in its response may use hedges when the information conveyed could have a certain degree of ambiguity. For instance, while telling about weather, the response were be framed with 'maybe cloudy' or 'looks like it would be sunny'.
- iii. For tasks which require a greater involvement of the user and may have a risk involved, asking for confirmations was recommended. Interrogating statements with the purpose of confirmations were used in contexts like navigation, schedule management and online payments.
- iv. The participants preferred to stress upon various instances of the voice assistant's replies. When asked, the reason for putting stress was to bring emphasis of the listeners to new information added by the voice assistant. For instance, Names, time values, location names and other nouns were stressed upon. Also if there was an action which maybe expected on users behalf was seen to be conveyed using stress in voice.
- v. Interjections were required to be highly context specific. If a user query was followed by an interjection, the user gets an impression of his query being understood well. Only when the context of the user's query was perceived as a positive or pertaining to a task which involves low risks, would a voice assistant react with an excited interjection like "ahaan" or "wow!" However when the context is serious in nature, in general the use of interjections was lower in number.

As the next step, few of these raw dialogue generated by the participants in this co-creation session were then given to 6 animation artists and designers with the intent of generating the visuals of the gestural aspect of these communication. We were discovering various gestural cues like smiles, frowns, eyebrow shapes, head nods, head positions, body postures, etc. [14, 15]. Table 2 presents some of the visual responses from the participants and some of the analysis drawn from those in terms of guidelines for gestural expression.

Table 2. Visual responses from the participants and analysis on gestural expressions

	Voice Assistant Natural Response			
	<i>Oh! What happened!</i>	<i>I'm afraid Dr. Johnson is not at his clinic</i>	<i>but he is at the hospital</i>	<i>Should I make an appointment for you?</i>
Eyes	Very wide open	Dim	Slight gaze aversion (indicating elsewhere)	Wide open, Staring
Eyebrows	Raised	Raised, shranked	Normal, straight	One pointing up other pointing down, stretched
Head position	Stable, forward	Slightly down	Moderate nodding, straight	Stable, straight
Expressions	Disappointment, sadness	Anxious, helplessness	Thinking	Concerned, showing sub-ordination
Body gesture/ Others	Dropped shoulders, slightly bent forward	Raised shoulders, raised hands	Finger pointing out	Bent knees, highly bent forward
Participant Responses				

6 Conclusions

The underlying goal of our study was to devise a methodology for deriving the appropriate behavior of a voice assistant avatar on mobile devices. Where many researchers have tried to formulate various methodologies for the same, ours was a user centered design approach. After carrying out the various activities with keeping user empathy in mind, we were able to come up with a few design directions for approaching the challenge of creating an avatar based voice assistant. Through the three phases of study, we identified few insights on the social behavior of the voice assistant, evolved the appropriate personality of the voice assistant as desired by the users and also formulated few guidelines for linguistic, speech as well as non-verbal expression constructs to be used by the voice assistant.

In summary, we identified that behavior of a voice assistant should be a moderate one, under various scenarios; the expression can vary up to only a certain range. User and assistant’s mood space should be as congruent as possible for being a reactive companion but not an over-reactive one. The voice assistant is expected to be very proactive and must interrupt the user immediately if the matter is more urgent and important than current task. Voice Assistant must avoid all updates from outside and become a single accessible point for all updates at a later stage. Preferred time to give updates/notifications is when user is switching from an ongoing task to another. From the personality aspect we found that the overall results from the personality elicitation exercise was quite similar to the actual behavior exhibited by the human personal assistants in the workplace. Voice assistant was expected not to be over sympathetic; and inventiveness comes out to be a much desired one.

In the dialogue co-creation activities, a preference in the informality of language was primarily noted. From the responses of the participants in this activity we were able to formulate few guidelines to design natural responses for the voice assistant with use of various languages and linguistic constructs like hedges, discourse makers, intonations, pauses, stresses, etc. For instance, interjections should be used with newly introduced contexts, must immediately follow user's query and must be uttered with high speech rate. It should mostly be avoided for ongoing dialogue context. Pauses are expected before giving alternatives, suggestions or making any enquiry from the user. Speech Rate could be fast for commonly used statements, but slower while asking any user's decision or updating any information such as date, time, status, etc.

Finally, with all the design guidelines, dialogues and personality definitions evolved from our research we designed the prototype of the natural voice assistant avatar. Going forward, we plan to extend this work with participants from few other geographical origins to consider cultural sensitivity in the design. Next goal would also be to perform experimental evaluation of the voice assistant prototype with the users.

References

1. Breese, J., Ball, G.: Modeling emotional state and personality for conversational agents. Rapport technique MSR-TR-98-41, Microsoft research (1998)
2. Norman, D.: How might humans interact with robots. Keynote address to the DARPA-NSF Workshop on Human-Robot Interaction, San Luis Obispo, CA (2001)
3. Ball, G., Breese, J.: Emotion and personality in a conversational agent. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.) *Embodied Conversational Agents*, pp. 189–219. MIT Press, Cambridge (2000)
4. Torrey, C., Fussell, S.R., Kiesler, S.: How a robot should give advice. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 275–282. IEEE (2013)
5. Meerbeek, B., Hoonhout, J., Bingley, P., Terken, J.: Investigating the relationship between the personality of a robotic TV assistant and the level of user control. In: *The 15th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN 2006*, pp. 404–410. IEEE (2006)
6. Heylighen, F., Dewaele, J.-M.: Formality of language: definition, measurement and behavioral determinants. *Interne Bericht, Center "Leo Apostel", Vrije Universiteit Brussel* (1999)
7. Becker, C., Kopp, S., Wachsmuth, I.: Why emotions should be integrated into conversational agents. In: *Conversational Informatics: An Engineering Approach*, pp. 49–68 (2007)
8. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161 (1980)
9. Dryer, D.C.: Getting personal with computers: how to design personalities for agents. *Appl. Artif. Intell.* **13**(3), 273–295 (1999)
10. Costa, P.T., MacCrae, R.R.: Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual. *Psychological Assessment Resources* (1992)
11. Pennebaker, J.W., King, L.A.: Linguistic styles: language use as an individual difference. *J. Pers. Soc. Psychol.* **77**(6), 1296 (1999)

12. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. *Speech Commun.* **40**(1), 5–32 (2003)
13. Brown, P.: *Politeness: Some Universals in Language Usage*, vol. 4. Cambridge University Press, Cambridge (1987)
14. Cassell, J., Vilhjálmsson, H.: Fully embodied conversational avatars: making communicative behaviors autonomous. *Auton. Agent. Multi-Agent Syst.* **2**(1), 45–64 (1999)
15. Ball, G., Breese, J.: Relating personality and behavior: posture and gestures. In: Paiva, Ana C.R. (ed.) *IWAI 1999. LNCS*, vol. 1814, pp. 196–203. Springer, Heidelberg (2000)