A vision and speech enabled, customizable, virtual assistant for smart environments

Giancarlo Iannizzotto
Dept. for Cognitive Sciences, Psychology, Education and Cultural Studies (COSPECS)
University of Messina
Messina, Italy
ianni@unime.it

Lucia Lo Bello
Department of Electrical, Electronic and Computer Engineering (DIEEI)
University of Catania
Catania, Italy
lucia.lobello@unict.it

Andrea Nucita
Dept. for Cognitive Sciences, Psychology, Education and Cultural Studies (COSPECS)
University of Messina
Messina, Italy
anucita@unime.it

Giorgio Mario Grasso
Dept. for Cognitive Sciences, Psychology, Education and Cultural Studies (COSPECS)
University of Messina
Messina, Italy
gmgrasso@unime.it

***Abstract*—Recent developments in smart assistants and smart home automation are lately attracting the interest and curiosity of consumers and researchers. Speech enabled virtual assistants (often named smart speakers) offer a wide variety of network-oriented services and, in some cases, can connect to smart environments, thus enhancing them with new and effective user interfaces. However, such devices also reveal new needs and some weaknesses. In particular, they represent faceless and blind assistants, unable to show a face, and therefore an emotion, and unable to 'see' the user. As a consequence, the interaction is impaired and, in some cases, ineffective. Moreover, most of those devices heavily rely on cloud-based services, thus transmitting potentially sensitive data to remote servers. To overcome such issues, in this paper we combine some of the most advanced techniques in computer vision, deep learning, speech generation and recognition, and artificial intelligence, into a virtual assistant architecture for smart home automation systems. The proposed assistant is effective and resource-efficient, interactive and customizable, and the realized prototype runs on a low-cost, small-sized, Raspberry PI 3 device. For testing purposes, the system was integrated with an open source home automation environment and ran for several days, while people were encouraged to interact with it, and proved to be accurate, reliable and appealing.**

***Index Terms*—Smart home, virtual assistant, computer vision, deep learning.**

I. Introduction

In the last decade the concept of smart assistant has become widely known and gained large popularity. Commercial devices such as Amazon Alexa, Google Home, Mycroft are able to interact with the user by means of speech recognition and speech synthesis, offer several network-based services and can interface with smart home automation systems, enhancing them with an advanced user interface. The diffusion of such speech-enabled smart assistants is constantly spreading, mainly thanks to the availability of a large number of network services and of an increasing number of additional skills, or capabilities, that can be easily added to the smart assistants. However, their potential is still limited by their inability to acquire real-time visual information from video data, either about the user or the environment. This also poses some critical security issues due to the fact that most speech-enabled smart assistants do not support effective authentication mechanisms, while being able to trigger security- critical actions. Face recognition or other identification mechanisms should be required before accepting voice commands, for such devices [1].

Current smart assistants can talk and listen to their users, but cannot "see" them. Moreover, in most cases they do not feature any kind of visual emotional feedback. They are blind and faceless to the user, thus their interaction is often impaired and incomplete and, therefore, less effective and efficient [2].

To overcome the problem listed above, this paper introduces an architecture for building vision-enabled smart assistants, provided with expressive and animated graphical characters and speech recognition and synthesis. The proposed architecture is specifically devised for, but not limited to, interfacing with smart home and home automation platforms. The resulting smart assistant aims at engaging the user in a very involving and effective interaction, exploiting multimodal and nonverbal communication.

In the next sections, this paper reports a brief description of the related work (Sect. II), a description of the proposed architecture and the developed prototype (Sect. III) and the produced preliminary experimental results (Sect. IV). Final remarks and future plans conclude the paper (Sect. V).

II. Related Work

Initial research on the effects of embodied virtual agents on human-computer interaction and on the 'persona effect', i.e., the positive effect of the presence of a lifelike character in an interaction environment, dates back more than 20 years [3]. Since then, the original findings have been confirmed several times and in several different applications, while the relevant literature grew enormously, covering a large number of technologies, applications and approaches [4] [5] [6]. Probably one of the most advanced, and recent, embodied virtual agents is SARA, described in [7], which features very accurate and complex abilities for affective and expressive human-computer interaction. However, SARA is principally aimed at affective interaction, by analyzing the voice and the

978-1-5386-5024-0/18/$31.00 ©2018 IEEE

facial expressions of the user as well as her voice intonation and the spoken text. Moreover, SARA is a large, complex system that could be hardly squeezed into a cheap, small-sized and resource-constrained hardware. Several other architectures and implementations were proposed in the literature [8] [9], however, in our knowledge, so far no embodied, vision- and speech- enabled virtual agents have been presented and released to the public, that are able to recognize the users discriminating their faces and to run on inexpensive and small sized consumer devices.

The lack of face identification abilities of most virtual agent software is probably due to the shortage, in the past, of suitable lightweight, yet effective and accurate, face identification approaches. In most cases, until a few years ago, face recognition was inaccurate or required powerful computational resources for the recognition process or for the off-line enrollment of the user pictures [10]. The recent application of deep neural networks is producing a disruptive change in this trend, allowing the development of effective, accurate and lightweight techniques for face recognition [11], that are currently also exploited for user identification in smartphones. It is about time to integrate those technologies into a virtual assistant.

III. System Architecture

The proposed architecture is fully modular. It is composed of a set of services, a graphical frontend and a coordinator that leverages on the services to offer to the user a multimodal and involving interaction with the connected home automation and smart assistant systems. In general, each service corresponds to a class of service modules, all offering the same service and exposing the same interface, but characterized by different performance, computational intensity, memory footprints, degree of portability (some modules may rely on proprietary services) and by their potential dependence on external (cloud) services. As a consequence, according to the specific requirements of the smart environment and user needs, a customized smart assistant can be built by combining a suitable set of modules.

The modules communicate through sockets and RESTFul connections, so, if needed, different modules can be allocated on different processing nodes. In principle, each module might be allocated on a separate node, however, most developed modules have limited requirements on computational power and memory resources. As a consequence, in most cases, a single node is sufficient to run all the involved modules.

An example of the described approach is the Text To Speech (TTS) service. In the proposed architecture, the associated class currently contains three different modules, called Flite2 module, MaryTTS module and GoogleTTS module. The MaryTTS module produces very natural utterances and supports multiple languages and voices. However, it has quite a large footprint (up to 1GB RAM), thus it is not suitable when other memory-intensive modules are involved. The Flite2 module produces natural, clear and understandable utterances, it is lightweight in terms of memory footprint and computational needs, but currently supports just a few languages and offers few voices (mostly male and US English or Indic speaking). The GoogleTTS module relies on the well-known Google Cloud Speech API [12]. Despite depending on cloud services and requiring a paid subscription, Google Cloud Speech API offers the benefits of supporting multiple languages and featuring both male and female, very natural, voices.

A. The Red Virtual Assistant

In order to demonstrate the proposed architecture, in this paper a specific virtual assistant is introduced, that exploits 6 service modules, a graphical frontend and a coordination module. The virtual assistant is named ‘Red’, after the name given to the frontend interactive character (a red fox). The structure of Red is shown in Fig. 1.

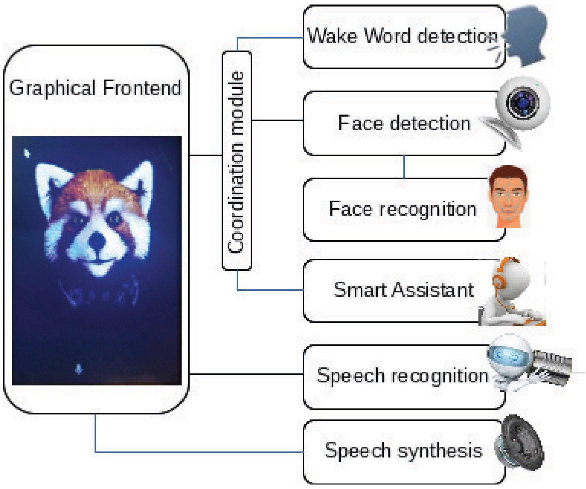

Fig. 1. Schematic representation of the modular structure of Red.

B. The Graphical Frontend and the Speech recognition module

The Graphical Frontend is based on an HTML5 document, containing the Javascript code needed to communicate through a Websocket connection with the Coordination module and through a RESTful API with the Speech synthesis and the Speech recognition modules. The character chosen for the interface, the Red fox, was animated in order to create ad hoc, expressive and meaningful videos, that are combined and synchronized with the speech at run time. This results in an involving and realistic interaction with the user (see Fig. 2).

The HTML5 document is delivered by an HTTPS server, locally installed, to a Chromium web browser, that was specifically chosen as it is open source and fully supports the Google Speech To Text (GSTT) service. The GSTT service was selected as it provides an effective and free solution for reliable and multilingual speech recognition. As a consequence, the Speech recognition module was realized by means of an external service. An alternative module based on the Mozilla open source implementation of DeepSpeech [13], that can run locally, is currently being investigated.

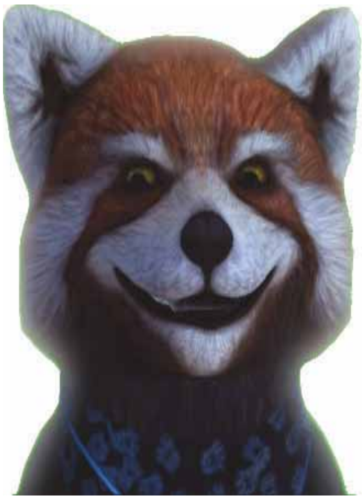

Fig. 2. The Red interactive character.

The Speech recognition module is normally idle, not recording nor sensing audio input, in order to preserve the user's privacy. It is explicitly activated by the Graphical Frontend in predetermined phases of the interaction (e.g., after posing a question to the user) and when the Coordinator module signals that the Wake Word has been detected by the Wake Word module.

C. The Speech synthesis module

The Speech synthesis service provides the speech ability to the Graphical Frontend. Two Speech synthesis modules were taken into consideration for Red: MaryTTS, based on the MaryTTS software [14], and the Flite2 module, a lightweight RESTful TTS server integrating the popular Flite open source TTS software [15] in its version 2.0. Due to the main memory limitations of the adopted hardware, featuring only 2GB of RAM, the lighter Flite2 module was chosen. Although during the tests Red interacted only in English language, an Italian voice developed for an earlier version of Flite [16] is currently being ported to this module. A version of the Flite2 module was also released as an open source node.js package[1].

D. The Wake Word detection module

The Wake Word detection service allows the user to trigger the Speech recognition service and give commands or ask questions to Red. A Wake Word module continuously scans the audio input and, when a predetermined utterance is detected (similarly to "Hey, Alexa" for Amazon Alexa, or "Hey, Google" for Google Assistant), it alerts the Coordination module. There are currently two modules in the Wake Word class, namely, the Snowboy module, integrating the SnowBoy deep neural networks-based software [17], and the PocketSphinx module, based on the homonymous open source software developed by Carnegie Mellon University [18]. The first module relies on a proprietary library containing the deep neural network that is the foundation of the Snowboy software, and thus is not fully portable. Conversely, the second module is open source and fully portable, but the procedure for encoding the wake word is cumbersome. Both modules were tested with Red, producing very similar performance using "Hey, Red!" as the Wake Word.

E. The Face detection module

The Face detection service allows the virtual assistant to detect the presence of a user in front of the device, thus contributing to enable the kind of interactivity that is totally missing in the most common virtual assistants. A Face detection module continuously scans the video input from a connected webcam and, whenever it detects a human face, it alerts the Coordination module. When a face is detected, Red turns its attention to the user facing the camera and greets her. If the identity of the user is available (as the user has been recognized by the Face recognition module), the user is called by name and every change in the identity of the user in front of the camera is signaled by a corresponding utterance ("Hi <name of the user>, how can I help you?").

The Face detection module chosen for Red is the Haar module, based on the fast and lightweight face detector by Viola and Jones [19] and optimized in the most recent version of the OpenCV library [20], but other modules were added to the Face Detection class to be tested with Red.

F. The Face recognition module

The Face recognition service allows the virtual assistant to recognize the user in front of the camera by matching her face against a set of photographs that reside in a local database. The number of photographs is not fixed and new ones can be added at any time, as the service protocol has a specific command for adding a new photograph to the collection, together with the associated identity. Although the service is not intended as a security tool, it has to be very accurate in order to keep the necessary reliability and the trust of the user. Moreover, a correct identification of the user is needed for maintaining the interaction context (the portion of the past interaction needed to correctly deal with the current dialogue). As a consequence, only the most accurate and recent approaches to face recognition were taken into account for the implementation of the Face recognition modules. The technique adopted for the module used for Red is based on an image metric realized through a ResNet-34 deep neural network [21], as modified and trained by Davis King [22] for the Dlib library [23]. This technique approaches an accuracy of 99.38% on the standard "Labeled Faces in the Wild" benchmark [24].

G. The Smart assistant module

A smart assistant is a software agent that can perform tasks or services for human users. Modern smart assistants rely on natural language processing (NLP) for parsing the input from the user and evincing commands and requests, and, in some cases, exploit artificial intelligence to elaborate articulated and

[1] npm package named ianniTTS (https://www.npmjs.com/package/iannitts)

meaningful answers when needed. The user input is usually textual (e.g. chatbots), vocal (as in the case of modern smart speakers). In some cases, it can also use images, that can be submitted to the agent by taking pictures with a smartphone. The Smart assistant module in Red acts as an interface to a smart assistant platform, thus integrating its services into the Virtual Assistant. Currently the class only contains the Mycroft module, that integrates the Mycroft open source smart assistant platform [25] into Red. Mycroft provides basic services and a minimal conversational ability, but it is modular and new services can be added by easily installing new "skills". Thanks to the Mycroft module, Red can answer to questions about the weather or the time worldwide, retrieve information from a web search engine or from Wikipedia, directly turn on a lamp or interface to a home automation platform, leave a message for another user, set an alarm, and more.

IV. Experimental results

The development, and, as a consequence, the experimentation of the Red virtual assistant, are still ongoing work. In order to produce a preliminary evaluation of both Red performance and user experience, three identical prototypes were deployed, all running the same software, in a students lab. Each prototype ran on a Raspberry PI 3 mod B card, equipped with an USB webcam with a microphone, a small speaker and a 3.5" color LCD display (see Fig. 3). In the lab, each prototype was positioned on a different desk, facing the user's chair, slightly to the side.

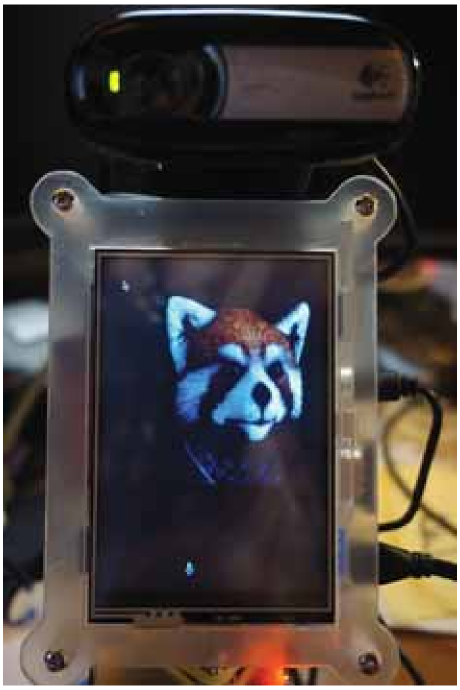

Fig. 3. A picture of one of the three Red prototypes.

The smart assistant architecture has a modular and customizable graphical interface, based on HTML5 documents and animations. Specifically for the Red prototype, 5 animations were produced, each one representing a different facial expression:

- Vacant expression, shown when no user is detected in front of the camera. The agent shows curiosity for the environment, looks around, searches for someone.
- Surprised expression, shown when a user, different from the last one seen before, is recognized in front of the camera. The user is also greeted with her name. If the user is unknown, i.e., Red does not have a picture of her in its database, the user is named "Stranger".
- "No no" expression, used to reject undesirable interactions (such as those containing contemptible language).
- Silent expression, adopted when some user is detected in front of the camera. Shows attention to the user.
- Chatting expression, used while Red is speaking. This expression is realistic and synchronized with the speech.

The users were given 30 minutes for getting accustomed to the presence of the virtual agent on their desk, before the real tests began. After a few minutes, depending on their age and curiosity, the users stopped being distracted by the device, that had gradually become a part of the environment. However, as soon as a chance to ask for its services appeared, the users never hesitated to bring it into play.

The actual test session was composed of a set of 8 tasks, involving different services and interactions. A total of 15 subjects (9 women and 6 men) were involved in the experiments. The subjects were 10 students, 18-24 years old, not engaged in technology-oriented studies, three young researchers, 25-28 years old, and two professors, 40 and 50 years old. Each user was asked to perform the whole set of tasks once, so each task was performed 15 times over the whole experiment.

For each task, measures were acquired regarding the number of fully successfully tasks, the number of partial failures (attempts that needed to repeat the interaction at least once to get the task successfully completed) and the number of full failures (attempts that produced a wrong result). The measures are reported in Table I.

TABLE I
Successfulness of the task, in percentage (session I). 15 trials for each task were performed.

Task	***Successful***	***Partial Failure***	***Failure***
Face detection	100%	-	-
Face recognition	100%	-	-
Ask for the time in New York	100%	-	-
Ask for the weather in Sidney	100%	-	-
Ask about Arthur C. Clarke	93.33%	6.67%	-
Switch on the light on the desk	100%	-	-
Ask who was sitting at the desk 10 minutes before	100%	-	-
Set an alarm in 10 minutes	93.33%	6.67%	-

Given the low number of participants to the tests, the very high score in face recognition was easily expected. A more realistic figure for the face recognition approach is provided in [22]. Nevertheless, this is indeed a typical scenario for home automation, where the number of potential users is, in most cases, that of the family components. The partial failures for the cases “Ask about Arthur C. Clarke” and “Set an alarm in 10 minutes” (in both cases 1 partial failure out of the whole set of tests) might have been partly originated by the low quality of the audio input. The adopted hardware platform does not provide high quality audio input or noise canceling. Most commercial devices, such as the Amazon Alexa, feature noise cancelling microphone arrays, but such a hardware is expensive and it was not considered necessary for our setup, as the user is supposed to be sitting in front of the device during interaction.

As the virtual assistant is speech-enabled, it was decided to make it directly pose the evaluation questions to the users. Red asked a number of question and automatically recorded the answers of each user, in textual form, on an Excel file. Although making the subject of the evaluation ask questions about its own performance might produce biased results, this was considered as part of the experiment, aimed at further revealing the degree of involvement of the users with the “subject” of the evaluation.

The users showed strong interest in this approach and answered the questions with attention and accuracy. Table II reports the results of the interviews.

The users were Italian speakers, while the interaction with Red was entirely in English. As a consequence, the small fraction of partial failures, also pointed out in the responses to the user experience questionnaires, might also be originated by an imperfect pronunciation.

In order to attain a high realism, and therefore a better “persona effect”, the assistant face is always slightly moving. This was in some cases considered “a bit creepy”, and somehow distracting. Although occurring very rarely, this might be a case of trespassing to the “uncanny valley” [26], despite the fact that the chosen character is not human.

TABLE II
USER EXPERIENCE QUESTIONNAIRE (SESSION I). ANSWERS REPORTED IN PERCENTAGE

Task	***Yes***	***More Yes than No***	***More No than Yes***	***No***
Did you enjoy the overall experience?	100%	-	-	-
Did you clearly understand my spoken messages?	93.33%	6.67%	-	-
Did I promptly catch your commands?	93.33%	-	6.67%	-
Would you like to have me on your desk at home?	100%	-	-	-
Would you enjoy my services on a daily basis?	100%	-	-	-
Were you able to concentrate on your work while I was sitting on your desk?	86.67%	6.67%	6.67%	-

Overall, the experience was considered very positive and the users declared to be interested in continuing the experimentation with the virtual assistant.

In order to highlight the effects of character animation and user face detection and recognition on the user interaction, the experiments were repeated with a disembodied version of Red. The face detection and recognition modules and the graphical interface were deactivated and the device was positioned horizontally, in order to ‘hide” it in the environment. The users could still evoke the agent by calling it by name (“Hey, Red!”), ask their questions and get their responses, however, they would not be recognized as different persons and would not get a visual feedback from the agent.

For the sake of avoiding any interference with the previous testing session, a new group of 12 users was selected, 5 male and 7 female, in the range 19-25 years old. Again, an initial 30 minutes interval was granted to the users to get accustomed to the presence of the disembodied virtual agent, which would respond to its name and reply to direct questions. After that interval, the real tests were performed.

TABLE III
SUCCESSFULNESS OF THE TASK, IN PERCENTAGE (SESSION II). 12 TRIALS FOR EACH TASK WERE PERFORMED.

Task	***Successful***	***Partial Failure***	***Failure***
Ask for the time in New York	100%	-	-
Ask for the weather in Sidney	91.67%	8.33%	-
Ask about Arthur C. Clarke	100%	-	-
Switch on the light on the desk	100%	-	-
Set an alarm in 10 minutes	100%	-	-

The results obtained from this second experimental session are reported in Table III. Unsurprisingly, the relevant accuracy results did not deviate significantly from those in Table I, as nothing changed in the smart agent architecture. Indeed, an annotation was taken by the observers during this session, reporting that the users tended to speak more slowly and articulated their words more than during the first session. Besides the inability to recognize the user, and thus the inability to maintain a reliable context-aware interaction, no other significant differences were found between the two sessions. However, while comparing the interaction logs from the two experimental sessions, some more interesting details emerged: before the first session, during the initial 30 minutes interval, after the first few minutes needed to get accustomed to the presence of the virtual assistant, the users actually started “chatting” with the virtual assistant. A number of attempts were made to establish some kind of informal interaction, with questions such as “how old are you?” or “what are you?”, “where do you come from” and “do you like music”. Luckily enough, a very rudimentary approach to conversational interaction had already been added to the coordination module, so the final experience for the users was not too frustrating. Notably, the described attempts to informal interaction did not appear in the logs of the second experimental session. As the users were different from those of the first session, such attempts were instead expected. A plausible explanation is that the disembodied agent did not tempt the users into considering it as a potential conversation partner. This hypothesis brings both good and bad news. The good news is that the embodied agent actually succeeded in making the interaction significantly more natural and attractive. The bad news is that such a more natural interaction requires an effective natural language processing interface and a context-aware conversational agent [27] in order to fulfill the increased expectations of the users.

The importance of the information gathered from the session logs was confirmed by the results of the interviews with the users involved in the second experimental session. As reported in Table IV, the users were much less impressed by the disembodied agent than by the embodied agent in the first session. 5 users out of 12 declared that the agent was nothing new with respect to the virtual assistant in their smartphone, and that the latter was faster, even thought it could not turn a lamp on. Overall, the net outcome was that the ability to visually interact with the user is definitely a significant plus for a virtual agent.

TABLE IV
USER EXPERIENCE QUESTIONNAIRE (SESSION II). ANSWERS REPORTED IN PERCENTAGE

Task	***Yes***	***More Yes than No***	***More No than Yes***	***No***
Did you enjoy the overall experience?	33.33%	66.67%	-	-
Did you clearly understand my spoken messages?	91.66%	8.33%	-	-
Did I promptly catch your commands?	91.66%	-	8.33%	-
Would you like to have me on your desk at home?	33.33%	66.67%	-	-
Would you enjoy my services on a daily basis?	33.33%	66.67%	-	-
Were you able to concentrate on your work while I was sitting on your desk?	100%	-	-	-

V. CONCLUSIONS AND FUTURE WORK

In this paper a software architecture for building lightweight, vision and speech-enabled virtual assistants for smart home and automation applications was presented. A complete prototype application was also developed, featuring a realistic graphic assistant able to show facial expressions and enabled with speech synthesis and recognition, face detection and face recognition for user identification. The assistant was also connected to a smart home assistant platform, thus building a complete “embodied” virtual home assistant that, differently from most common smart speakers, is able to “see” and “be seen” by the user and engage her in a multimodal interaction.

An explorative experimentation was carried out and was reported in the paper. The experimental results are satisfactory both in terms of accuracy and reliability, and in terms of user experience. In particular, the users appreciated the experience and the feeling with the interface and expressed their willingness to continue working with it. A simple counter-experiment was set up with a disembodied version of the virtual agent, showing that the ability to detect and recognize the user, as well as the graphical interface, largely improve the user experience.

The presented architecture is still under development and several improvements are being added. In particular, the virtual assistant ability to acquire visual information on the user and its surroundings also paves the way for further important applications, such as fall detection [28] or anomalous or dangerous behavior [29]. Each functionality will be implemented through a dedicated service module, possibly running on a separate node and connected with the others in real-time through an adequate wireless communication protocol [30], able to offer reduced energy consumption and packet loss rate [31].

REFERENCES

[1] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, “The insecurity of home digital voice assistants - amazon alexa as a case study,” *CoRR*, vol. abs/1712.03327, 2017.

[2] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, “Creating rapport with virtual agents,” in *Intelligent Virtual Agents* (C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, eds.), (Berlin, Heidelberg), pp. 125–138, Springer Berlin Heidelberg, 2007.

[3] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal, “The persona effect: Affective impact of animated pedagogical agents,” in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI ’97, (New York, NY, USA), pp. 359–366, ACM, 1997.

[4] *Embodied Conversational Agents*. Cambridge, MA, USA: MIT Press, 2000.

[5] E. André and C. Pelachaud, *Interacting with Embodied Conversational Agents*, pp. 123–149. Boston, MA: Springer US, 2010.

[6] B. Weiss, I. Wechsung, C. Kühnel, and S. Möller, “Evaluating embodied conversational agents in multimodal interfaces,” *Computational Cognitive Science*, vol. 1, p. 6, Aug 2015.

[7] Y. Matsuyama, A. Bhardwaj, R. Zhao, O. Romeo, S. Akoju, and J. Cassell, “Socially-aware animated intelligent personal assistant agent,” in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 224–227, Association for Computational Linguistics, 2016.

[8] M. Schroeder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wllmer, “Building autonomous sensitive artificial listeners,” *IEEE transactions on affective computing*, vol. 3, pp. 165–183, 4 2012. eemcs-eprint-22932.

[9] B. Martinez and M. F. Valstar, *Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition*, pp. 63–100. Cham: Springer International Publishing, 2016.

[10] F. Battaglia, G. Iannizzotto, and L. Lo Bello, “A person authentication system based on rfid tags and a cascade of face recognition algorithms,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, pp. 1676–1690, Aug 2017.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, June 2015.

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” *ArXiv e-prints*, Dec. 2017.

[13] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” *CoRR*, vol. abs/1412.5567, 2014.

[14] M. Schröder and J. Trouvain, “The german text-to-speech synthesis system mary: A tool for research, development and teaching,” *International Journal of Speech Technology*, vol. 6, pp. 365–377, Oct 2003.

[15] A. W. Black and K. A. Lenzo, “Flite: a small fast run-time synthesis engine,” in *4th ITRW on Speech Synthesis, Perthshire, Scotland, UK, August 29 - September 1, 2001*, p. 204, 2001.

[16] P. Cosi, F. Tesser, R. Gretter, C. Avesani, and M. Macon, “Festival speaks italian!,” in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001* (P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, eds.), pp. 509–512, ISCA, 2001.

[17] KITT.AI, “Snowboy hotword detection.” https://github.com/kitt-ai/snowboy, 2018.

[18] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, “Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, May 2006.

[19] P. Viola and M. J. Jones, “Robust real-time face detection,” *Int. J. Comput. Vision*, vol. 57, pp. 137–154, May 2004.

[20] Itseez, “Open source computer vision library.” https://github.com/itseez/opencv, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

[22] D. E. King, “High quality face recognition with deep metric learning.” http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html, 2017.

[23] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[24] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, *Labeled Faces in the Wild: A Survey*, pp. 189–248. Cham: Springer International Publishing, 2016.

[25] MycroftAI, “Mycroft, an open source artificial intelligence for everyone.” https://github.com/MycroftAI/mycroft-core, 2018.

[26] M. Mori, “Bukimi no tani [The uncanny valley],” *Energy*, vol. 7, no. 4, pp. 33–35, 1970.

[27] M. Jain, R. Kota, P. Kumar, and S. N. Patel, “Convey: Exploring the use of a context view for chatbots,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, (New York, NY, USA), pp. 468:1–468:6, ACM, 2018.

[28] F. Cardile, G. Iannizzotto, and F. La Rosa, “A vision-based system for elderly patients monitoring,” in *3rd International Conference on Human System Interaction*, pp. 195–202, May 2010.

[29] G. Iannizzotto and L. Lo Bello, “A multilevel modeling approach for online learning and classification of complex trajectories for video surveillance,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, 08/2014 2014.

[30] G. Iannizzotto, F. La Rosa, and L. Lo Bello, “A wireless sensor network for distributed autonomous traffc monitoring,” in *3rd International Conference on Human System Interaction*, pp. 612–619, May 2010.

[31] E. Toscano and L. Lo Bello, “A topology management protocol with bounded delay for wireless sensor networks,” in *2008 IEEE International Conference on Emerging Technologies and Factory Automation*, pp. 942–951, Sept 2008.