# Chapter 2
# The Dawn of the Conversational Interface

*Good morning, Theodore. You have a meeting in five minutes.*
*Do you want to try getting out of bed?*

Samantha in the movie Her: Official Trailer

**Abstract** With a conversational interface, people can speak to their smartphones and other smart devices in a natural way in order to obtain information, access Web services, issue commands, and engage in general chat. This chapter presents some examples of conversational interfaces and reviews technological advances that have made conversational interfaces possible. Following this, there is an overview of the technologies that make up a conversational interface.

## 2.1 Introduction

In the movie *Her* (2013), Theodore Twombly acquires Samantha, described as "the world's first intelligent operating system." Samantha is constantly available, just like a personal assistant, not only monitoring Theodore's calendar and answering his questions but also providing guidance and support in his personal life. They develop a relationship, and Theodore confesses to a friend that he is "dating" Samantha.

In the real world, being able to talk with a personal assistant in this way on a smartphone or other smart device has almost become a reality. Personal assistants, known by various names such as virtual personal assistants (VPAs), intelligent personal assistants, digital personal assistants, mobile assistants, or voice assistants, have become mainstream. Examples include Apple's Siri, Google Now, Microsoft Cortana, Amazon Alexa, Samsung S Voice, Facebook's M, and Nuance Dragon. Indeed, a search for "personal assistants" on Google Play toward the end of December 2015 returned 100 entries. Many of these VPAs help users to perform a variety of tasks on their smartphones, such as obtaining information using voice search, finding local restaurants, getting directions, setting the alarm, updating the calendar, and engaging in general conversation. Others provide more specialized functions, such as fitness monitoring, personalized preparation of drinks, and recipe planning.

We use the term *conversational interface* to refer to the technology that supports conversational interaction with these VPAs by means of speech and other modalities. To set the scene, we begin with some examples of the sorts of interactions that can be performed using such conversational interfaces.

**Fig. 2.1** Google Search box on a Nexus 5 smartphone. Google and the Google logo are registered trademarks of Google Inc., used with permission

## 2.2  Interacting with a Conversational Interface

The following examples are taken from interactions with the Google Now personal assistant, which is available for Android devices as well as for iPhones and iPads. Google Now can be activated by tapping on the microphone icon in the Google Search box, as shown in Fig. 2.1 and also on more recent Android devices (Android 4.4 onwards) by saying "OK Google". This activates speech recognition, leading to a screen displaying the instruction say "Ok Google".

There is a Google support page that provides information on how to turn on "OK Google" voice search along with examples of what you can say in a large variety of languages.[1] The following example shows the output from a query about the weather.

> User (spoken input): What's the weather in Belfast?
> Google Now (spoken output): It's 7 degrees and partly cloudy in Belfast.

In addition to the spoken response, there is also a visual display of the recognized question, a visual representation of the weather forecast, and, on scrolling down, the addresses of some relevant Web pages (Fig. 2.2).

The next example is a general knowledge query.

> User (spoken input): When was Belfast City Hall built?
> Google Now (spoken output): Belfast City Hall is 110 years old.

The response from Google Now displays the recognized question, the answer, a picture showing Belfast City Hall, when it was built, a map of its location, and, on scrolling down, links to some additional information (Fig. 2.3).

Our final example illustrates access to operations on the device, in this case to set an alarm.

---

[1]https://support.google.com/websearch/answer/2940021?hl=en. Accessed February 19, 2016.

**Fig. 2.2** Response to the query "What's the weather in Belfast" to Google Now on a Nexus 5 at 15:25 on February 10, 2016. Google and the Google logo are registered trademarks of Google Inc., used with permission
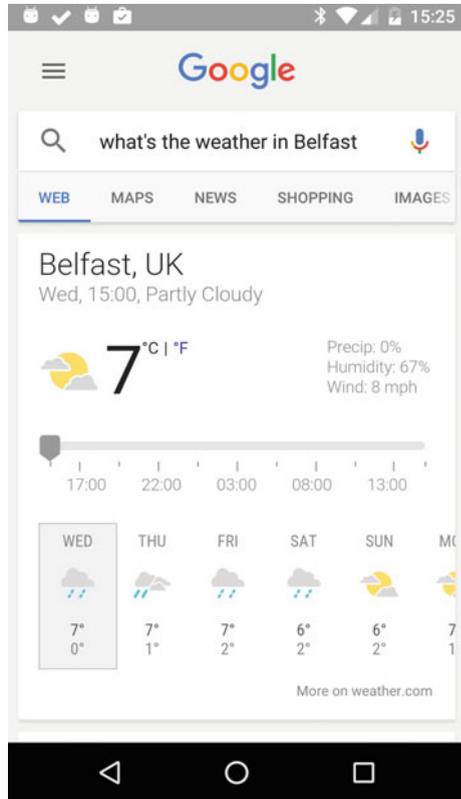
User (spoken input): Set the alarm for 9 o'clock tomorrow morning;
Google (spoken output): OK, 9 am, setting your alarm.

In addition to the spoken response, Google Now also presents the information shown in Fig. 2.4, which displays the recognized question, and a message confirming that the alarm has been set.

A wide range of device operations can be performed using voice, such as placing a call to a contact, sending a text, or launching an app. In many cases, the use of voice commands enables users to carry out these operations in fewer steps compared with traditional input methods. For example, the following steps would be required to set an alarm manually on a Nexus 5:

1. Tap on the clock icon (e.g. from the home screen).
2. Find the icon representing the alarm and tap on it.
3. Tap on the time displayed.
4. Adjust the hours and minutes to the required time.
5. Tap on "Done" to finish.

**Fig. 2.3** Response to the
question "When was Belfast
City Hall built" to Google
Now on a Nexus 5 at 15:58 on
February 10, 2016. Google
and the Google logo are
registered trademarks of
Google Inc., used with
permission
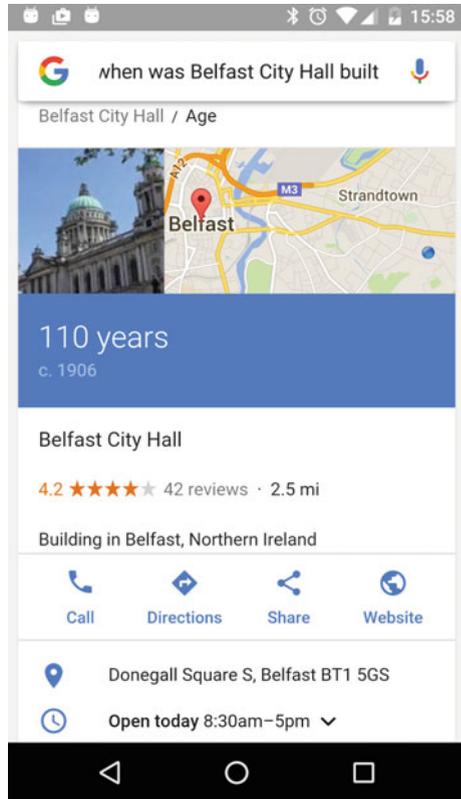


**Fig. 2.4** Response to the
command "Set the alarm for 9
o'clock tomorrow morning"
to Google Now on a Nexus 5
on February 10, 2016. Google
and the Google logo are
registered trademarks of
Google Inc., used with
permission
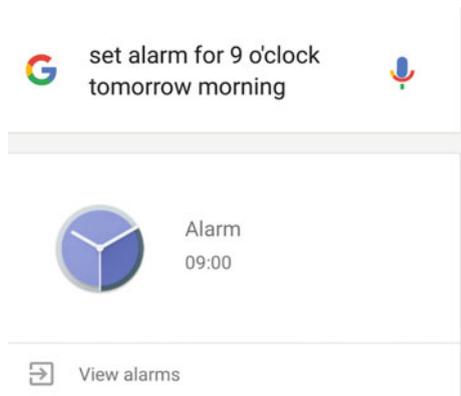
## 2.3   Conversational Interfaces for Smart Watches and Other Devices

Although conversational interfaces first appeared on smartphones, they are now also being deployed on various other devices such as smart watches, social robots, and devices such as Amazon Echo. In the future, we might expect conversational interfaces to be an integral part of the so-called Internet of Things (IoT), a massive network of connected objects, sensors, and devices that "talk" to each other and in some cases also communicate with humans.

A smart watch is a wearable device that provides many of the functions of a smartphone, such as notifications of incoming messages and emails, navigation, and voice search. Some smart watches also provide fitness tracking and heart rate monitoring. Users interact with a smart watch by speaking, by tapping on items on the screen, and by swiping the displayed cards. Some smart watches also have speakers that provide spoken output. Smart watches do not usually have an Internet connection so they have to be connected to a smartphone via Bluetooth pairing using software such as Android Wear—a dedicated software development kit (SDK) for wearables that is compatible with Android as well as iOS devices. One of the advantages of a smart watch is that actions such as responding to an email or text message can be performed on the fly using voice without having to take the phone from a pocket or handbag, find the required app, and tap in a reply.

Social robots allow users to perform tasks similar to those provided by a virtual personal assistant on a smartphone. However, because of their physical embodiment and because they often possess social qualities such as the ability to recognize and display emotions, they also provide social functions such as companionship for the elderly.

Conversational interfaces on devices such as smart watches and social robots provide many of the same capabilities that are already available on smartphones, although there may be differences in terms of the type of interface they provide. For example, the display on a smart watch is smaller than that on a typical smartphone, while some devices do not have a display for information but rely solely on voice for input and output. Chapter 13 discusses various devices, wearables, and robots; Chaps. 14 and 15 explore emotion and personality in conversational interfaces; and Chap. 16 provides a tutorial on how to develop apps for smart watches using the Android Wear SDK.

## 2.4   Explaining the Rise of the Conversational Interface

The conversational interface has for a long time been a vision of researchers in speech technology and artificial intelligence (AI), but until recently, this vision has only been realized in science fiction books and movies, such as 2001: A Space Odyssey, Star Wars, and many others. In 1987, Apple released a concept video

depicting a device called the Knowledge Navigator, a software agent that provided services similar to those of a present-day VPA and that possessed advanced communication capabilities, including excellent text-to-speech and perfect speech understanding. In 2001, Berners-Lee and colleagues put forward a vision for a Semantic Web in which Semantic Web agents would perform tasks such as checking calendars, making appointments, and finding locations (Berners-Lee et al. 2001). But it was not until 2011 that these visions were realized with the launch of Siri, generally recognized as the first voice enabled VPA.

Siri and similar conversational systems have been made possible as a result of developments in technology and of increasing user acceptance and adoption, as explained further in the following sections.

## 2.4.1  Technological Developments

Various technological advances have contributed to the recent rise of conversational interfaces.

**The renaissance of Artificial Intelligence**. Since the mid-1950s, researchers in artificial intelligence (AI) have wrestled with the challenge of creating computers that are capable of intelligent behavior. AI has gone through cycles of euphoria and rejection with some initial successes followed by some spectacular failures. At first, it was believed that intelligent behavior could be reproduced using models of symbolic reasoning based on rules of formal logic. This was known as the *knowledge-based approach*, in which the focus was on problems that are difficult for humans but easy for computers—for example, decision-making and playing chess. Knowledge-based systems (also known as expert systems) were developed in the 1970s and 1980s to assist with decision-making in complex problems such as medical diagnosis, while IBM's chess playing computer Deep Blue defeated a human world champion in 1996. However, it became evident that various aspects of intelligent behavior that are easy for humans but difficult for computers, such as speech recognition and image recognition, could not be solved using these symbolic approaches but required processes such as the extraction of patterns from data and learning from experience. As a result, *subsymbolic approaches* using neural networks and statistical learning methods have come to dominate the field. Several factors have contributed to the recent success of subsymbolic approaches: developments in graphics processing units (GPUs) that have enabled the massive parallel computations required to run neural networks; the availability of vast amounts of data (known as *big data*) that enable AI systems to learn and become increasingly more intelligent; and the development of new algorithms (known as *deep learning*) that run on GPUs and process these vast amounts of data.[2] A sign of the promise of this new AI is that many major companies such as Google, Microsoft, Amazon,

---

[2]http://www.wired.com/2014/10/future-of-artificial-intelligence/. Accessed February 19, 2016.

Facebook, and Baidu—China's leading internet-search company—have recruited the world's leading experts in deep learning to support their research and development work in areas such as search, learning, natural language understanding, and personal assistant technology.

**Advances in language technologies**. Language technologies have benefitted from the new AI. Speech recognition accuracy has improved dramatically since around 2012 following the adoption of deep learning technologies. There have also been major advances in spoken language understanding. Machine learning approaches to dialog management have brought improved performance compared with traditional handcrafted approaches by enabling systems to learn optimal dialog strategies from data. Furthermore, grand challenges in various areas of speech and language technology, including speech recognition, text-to-speech synthesis, spoken dialog management, and natural language learning, have promoted the exploration and evaluation of different systems and techniques using shared tasks and data, leading to technological advances and wider cooperation within the research communities (Black and Eskenazi 2009).

**The emergence of the Semantic Web**. The vision of the Semantic Web is that all of the content on the Web should be structured and machine-readable, so that search using the traditional approach of keywords as input has been replaced by semantic search based on the meaning of the input. Semantically tagged pages marked up using encodings such as Resource Description Framework in Attributes (RDFa) and large structured knowledge bases such as Google's Knowledge Graph have enabled search engines to better interpret the semantics of a user's intent, to return structured answers to queries, and, for virtual personal assistants such as Google Now, to support a question/answer type of interaction. Examples of the more complex types of question that can now be answered by the Google app are described here.[3]

**Device technologies**. Smartphones and other intelligent devices have become more powerful than the large personal computers of only a few years ago. Indeed, in one comparison, it was stated that a single Apple iPhone5 has 2.7 times the processing power of a 1985 Cray-2 supercomputer.[4] Moreover, since smartphones have access to a wide range of contextual information, such as the user's location, time and date, contacts, and calendar, the integration of this contextual information into conversational interfaces enables VPAs to provide help and support that is relevant and personalized to the individual user.

**Increased connectivity**. Faster wireless speeds, the almost ubiquitous availability of WiFi, more powerful processors in mobile devices, and the advent of cloud computing mean that resource-intensive operations such as speech

---

[3]http://insidesearch.blogspot.com.es/2015/11/the-google-app-now-understands-you.html. Accessed February 19, 2016.
[4]http://www.phonearena.com/news/A-modern-smartphone-or-a-vintage-supercomputer-which-is-more-powerful_id57149. Accessed February 19, 2016.

recognition and search can be performed in the cloud using large banks of powerful computers.

**The interest of major technology companies in conversational interfaces**. While previously interest in conversational interfaces for VPAs was limited to relatively small niche companies and to enthusiastic evangelists of the AI dream, now many of the largest companies in the world are competing to create their own VPAs, for example, Apple's Siri, Google's Google Now, Amazon's Alexa, Microsoft's Cortana, Facebook's M, and Baidu's Duer. These VPAs enable companies to more accurately profile the users of their VPAs, enabling them to promote their e-commerce services and thus gain a competitive advantage.

Notwithstanding these advances, there is still more work to be done before conversational interfaces achieve a level of performance similar to that of humans. For example, in looking for a possible way forward, Moore (2013) suggests that it is necessary to go beyond the domain of speech technology and draw inspiration from other fields of research that inform communicative interaction, such as the neurobiology of living systems in general.

## 2.4.2  User Acceptance and Adoption

Even if a product is technologically advanced, it will not succeed unless it is accepted and adopted by users. Until recently, it seemed that users stopped using their VPAs after an initial stage of experimentation. In some cases, they encountered problems such as speech recognition errors and so reverted to more accustomed and more accurate modes of input. Some users found amusement by saying "silly" things to their VPA to see what sort of response they would get. Furthermore, the proliferation of so many virtual personal assistants makes it difficult to select and adopt one particular VPA for regular use.

Evaluations of VPAs have so far been fairly informal, taking the form either of showdowns or of surveys. In a showdown, a large bank of questions is submitted to selected VPAs and the responses are analyzed. In one showdown, Google Now was compared with Microsoft Cortana and Soundhound's VPA Hound,[5] while in another, Google Now was compared also with Microsoft Cortana as well as with Siri.[6]

Conversational interfaces are appealing to users who wish to engage with Web services when on the go. Given the processing power and speed of modern smartphones as well as ongoing Internet connectivity, users no longer need to be located at a desktop PC to search for information or access Web services. Also, with devices becoming smaller to aid portability input is easier using a conversational

---

[5]http://www.greenbot.com/article/2985727/google-apps/android-virtual-assistant-showdown-google-now-vs-cortana-vs-hound.html. Accessed February 19, 2016.

[6]https://www.stonetemple.com/great-knowledge-box-showdown/. Accessed February 19, 2016.

interface compared with tapping on the soft keyboards of smartphones. In any case, some devices will not have keyboards but only microphones for voice input. This is likely to be the case as more and more devices become linked in the Internet of Things, where many of the devices will rely exclusively on voice for input and output.

Young people are also more likely to use conversational interfaces. In a recent study of the use of voice search, it was reported that teenagers talk to their phones more than the average adult and more than half of teenagers between 13 and 18 use voice search daily.[7] Voice search is also widely used in China in VPAs such as Baidu's Duer as it is more difficult to input text in Chinese and so speech is a more convenient input mode.

### 2.4.3  Enterprise and Specialized VPAs

Enterprise and specialized VPAs provide assistance in specific domains and for specific users. Enterprise and specialized VPAs can assist professionals in their work—for example, helping doctors to manage their workload, schedules, messages, and calls, and to obtain up-to-date and reliable information to assist with diagnosis. They can also assist customers to get help and information about a company's products.

IBM Watson for oncology is an example of a specialized VPA that helps oncologists to make evidence-based treatment decisions based on an analysis of an individual patient's medical records and a search for treatment options in a vast corpus of information from journals, textbooks, and millions of pages of text.[8] The Ask Anna VPA, developed by Artificial Solutions[9] to provide help to customers searching for information about products on the IKEA Web site, is an example of a customer-facing VPA.[10] Other examples include JetStar's Ask Jess virtual assistant, developed on the Nuance Nina platform,[11] that answers customers' queries about bookings, baggage, and seating,[12] and Next IT's Alme, a multimodal, multichannel, and multilanguage platform for customer service in domains such as health care, travel, insurance, finance, and retail.[13]

---

[7]https://googleblog.blogspot.fr/2014/10/omg-mobile-voice-survey-reveals-teens.html.    Accessed February 19, 2016.

[8]http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html.  Accessed  February 19, 2016.

[9]http://www.artificial-solutions.com/. Accessed February19, 2016.

[10]https://www.chatbots.org/virtual_assistant/anna3/. Accessed February 19, 2016.

[11]http://www.nuance.com/company/news-room/press-releases/Jetstar.docx. Accessed February 19, 2016.

[12]http://www.jetstar.com/au/en/customer-service. Accessed February 19, 2016.

[13]http://www.nextit.com/alme/. Accessed February 19, 2016.

VPAs can provide a commercial advantage for companies in the generation of advertising revenues and referral fees by directing users to specific services and Web sites that have been "chosen" by the assistant. Furthermore, as Meisel (2013) points out, they can promote a company's brand and services in a similar way to the company's Web site, but with the added value of a more personalized and more enjoyable interaction.

### 2.4.4   The Cycle of Increasing Returns

It has been predicted in a number of studies that the global market for VPAs will increase dramatically in the next few years. One factor in addition to those discussed above is the so-called cycle of increasing returns. User acceptance and adoption interact with developments in technology to produce a cycle of increasing returns. As performance improves, more people will use conversational interfaces. With more usage, there will be more data that the systems can use to learn and improve. And the more they improve, the more people will want to use them. Given this cycle, it can be expected that conversational interfaces will see a large uptake for some time to come and that this uptake will be accompanied by enhanced functionalities and performance.

## 2.5   The Technologies that Make up a Conversational Interface

In this book, we describe the various technologies that make up a conversational interface and that enable users to engage in a conversation with a device using spoken language and other modalities. In Part 2, we will focus on spoken language technologies, as these are the core components of the majority of current conversational interfaces, while in Part 3, we will describe additional aspects of the input and output such as the recognition and display of emotion and personality.

Looking first at conversational interfaces that make use of spoken language technologies, Fig. 2.5 shows the typical components of such a system and the information flow between the components.

Typically, such a conversational interface operates as follows. On receiving spoken input from the user, the system has to:

- Recognize the words that were spoken by the user (speech recognition).
- Interpret the words, i.e., discover what the user meant and intended by speaking these words (spoken language understanding).
- Formulate a response, or if the message was unclear or incomplete, interact with user to seek clarification and elicit the required information (dialog management).
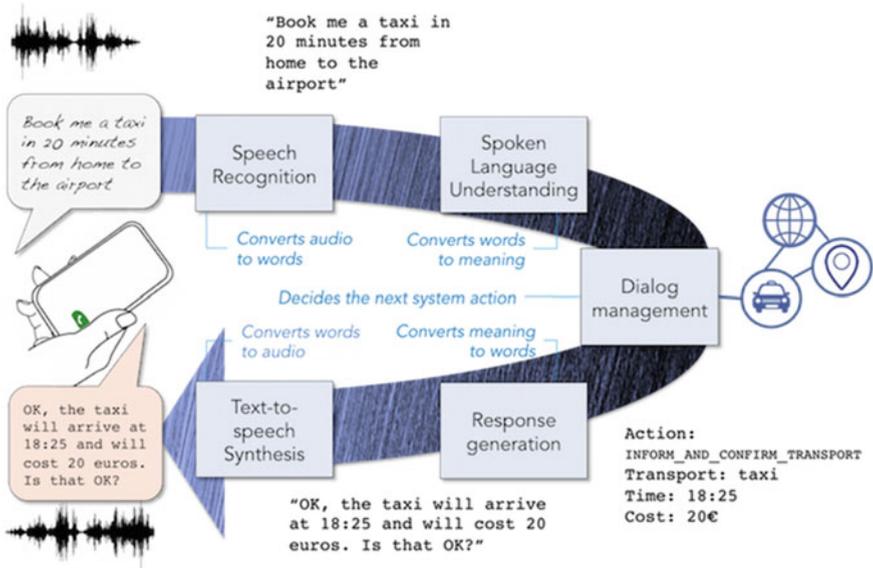
**Fig. 2.5** The components of a spoken language conversational interface

- Construct the response, which may be in the form of words or, as in the examples above, accompanied by visual and other types of information (response generation).
- Speak and display the response (text-to-speech synthesis).

However, when people engage in natural conversational interaction, they convey much more than just the meanings of the words spoken. Their speech also conveys their emotional state and aspects of their personality. Additionally, in face-to-face interaction, their nonverbal behaviors, such as their facial expression, gestures, and body posture, also convey meaning. Other information may also be transmitted when speaking to a conversational interface. For example, smartphones and other smart devices have built-in sensors and actuators that gather data about the user and the environment, including location, motion, orientation, and biosignals such as heart rate. Figure 2.6 shows these additional inputs to the conversational interface. We look at these in more detail in Part 3.

As shown in these figures, once the conversational interface has interpreted the input from the user, it constructs queries to Web services and knowledge sources in order to perform tasks and retrieve information to be output by the response generation component. Our focus in this book will be on the conversational interface, and we will not explore how these Web services and knowledge sources are accessed and how information to be presented to the user is retrieved, although in some of the laboratory sessions, we will show how to connect to some existing services such as Google Search and Google Maps.
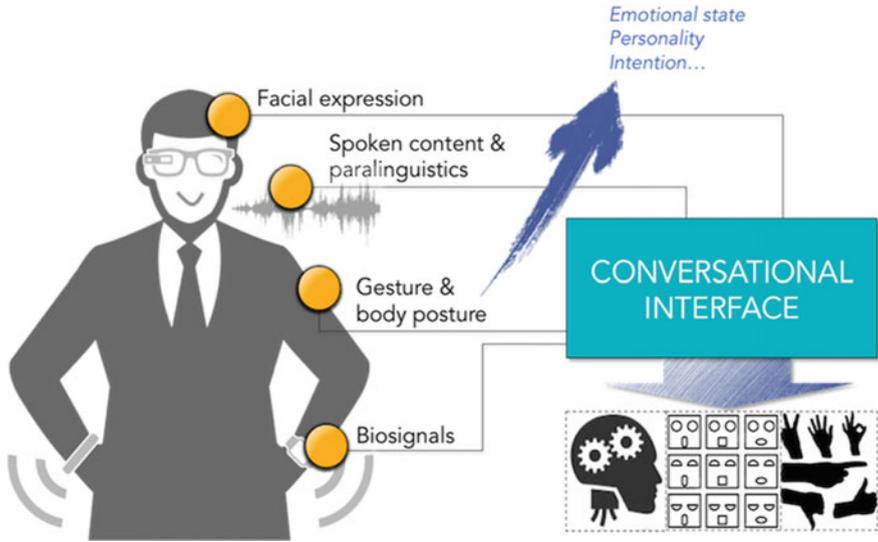
**Fig. 2.6** Additional inputs to the conversational interface

## 2.6   Summary

A conversational interface allows people to talk to their devices in an intuitive and natural way. In this chapter, we have presented some examples of conversational interfaces that are available on smartphones and other devices.

Conversational interfaces have been made possible by recent advances in technology, in particular:

- A renaissance in AI, in which deep learning has brought about dramatic improvements in speech recognition accuracy and more recently in spoken language understanding and dialog management.
- The development of powerful processors that support the massively parallel computations required for deep learning algorithms and that provide the processing power on small devices such as smartphones that were available only to supercomputers a few years ago.
- Advances in the technologies of the Semantic Web that enable almost instantaneous access to the vast stores of unstructured as well as structured knowledge on the Internet.

As a result of these technological advances, user acceptance of technologies such as the conversational interface has increased, leading to increased adoption and consequently producing more data from which systems can learn, in turn resulting in further improvements in the technology.

Before we go on to explore the various technologies that make up the conversational interface, we need to understand what we mean by "conversation." In the

next chapter, we look at the technology of conversation, in which we provide an overview of the contributions made by researchers in a variety of fields, including linguistics, psychology, sociology, and AI, to our understanding of the processes of conversational interaction.

**Further Reading**

There are a number of books and other resources that cover the topics of virtual personal assistants as well as related developments in AI and speech and language technologies. Trappl (2013) explores what is required for a personalized virtual butler and how such a virtual butler might be useful for an aging population. Stork (1998) is a collection of papers that examines the extent to which science fiction's HAL, the computer in the movie Space Odyssey 2011, has become technologically feasible—a bit dated now with recent advances in AI, but still worth reading. Meisel's online book The Software Society discusses the technology of virtual personal assistants,[14] while the short online book by Bouzid and Ma (2013) provides an introduction to the principles and techniques behind effective voice user interface design. An article by Ron Kaplan, Head of Nuance Communications' Natural Language Understanding R&D Lab, argues for conversational user interfaces as a replacement for the traditional graphical user interface.[15] The conversational interface is also the subject of a blog by John M Smart.[16] Pieraccini (2012) provides an accessible overview of speech technology and its history, and discusses the emergence of virtual personal assistants in the final chapter.

There are a number of blogs and other online resources that regularly discuss virtual personal assistants, including Amy Stapleton's Virtual Agent Chat,[17] Bill Meisel's Speech Strategy News,[18] and the Conversational Interaction Technology news board.[19] An online article by Buzzanga discusses recent advances in search that go beyond the use of keywords.[20]

**Exercises**

1. Watch the video "Behind the mic: the science of talking with computers."[21] This short video, produced by Google Creative Lab, features some of the key researchers in the field talking about speech recognition, language understanding, neural networks, and the use of speech to interact with smart devices.

---

[14]http://thesoftwaresociety.com/. Accessed February 19, 2016.

[15]http://www.wired.com/2013/03/conversational-user-interface/. Accessed February 19, 2016.

[16]http://www.accelerationwatch.com/lui.html. Accessed February 19, 2016.

[17]http://virtualagentchat.com/. Accessed February 19, 2016.

[18]http://www.tmaa.com/speechstrategynews.html. Accessed February 19, 2016.

[19]http://citia.lt-innovate.eu/. Accessed February 19, 2016.

[20]https://www.sla.org/wp-content/uploads/2015/06/2015_Buzzanga.pdf. Accessed February 19, 2016.

[21]https://youtu.be/yxxRAHVtafI. Accessed 2 March 2016.

2. Go to YouTube and search for "virtual personal assistants." You will find a large number of videos. Watch some videos that show examples of different VPAs and take note of the sorts of questions and commands that they can handle.

3. Look for videos on YouTube that show shootouts (or comparisons) of different VPAs, noting the questions and commands that are used in the tests. You can also find several interesting videos and demos created in the Interaction Lab of Heriot Watt University, Edinburgh.[22]

4. Go to Google Play[23] and search for "virtual personal assistants." A large number of VPAs will be returned. Select and download two that you can use for a comparison test. For example, you could download Assistant and Call Mom, as these are featured later in the book. Using the questions and commands noted in exercise 2, test and compare the two VPAs, noting where they succeed and where they fail. For inputs that fail, distinguish between failures due to speech recognition errors and failures due to the back-end components of the app (e.g., if it is not able to make a correct search for the item you requested).

# References

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284:34–43. doi:10.1038/scientificamerican0501-34

Black AW, Eskenazi M (2009) The spoken dialog challenge. In: Proceedings of the 10th annual meeting of the special interest group in discourse and dialog (SIGDIAL 2009), Queen Mary University of London, September 2009, pp 337–340. doi:10.3115/1708376.1708426

Bouzid A, Ma W (2013) Don't make me tap: a common sense approach to voice usability. ISBN 1-492195-1-7

Meisel W (2013) The personal-assistant model: unifying the technology experience. In: Neustein A, Markowitz JA (eds) Mobile speech and advanced natural language solutions. Springer, New York, pp 35–45. doi:10.1007/978-1-4614-6018-3_3

Moore RK (2013) Spoken language processing: where do we go from here? In: Trappl R (ed) Your virtual butler: the making of. Springer, Berlin, pp 119–133. doi:10.1007/978-3-642-37346-6_10

Pieraccini R (2012) The voice in the machine: building computers that understand speech. MIT Press, Cambridge

Stork D (ed) (1998) HAL's legacy: 2001's computer as dream and reality. MIT press, Cambridge

Trappl R (ed) (2013) Your virtual butler: the making of. Springer, Berlin. doi:10.1007/978-3-642-37346-6

---

[22]https://sites.google.com/site/hwinteractionlab/demos. Accessed March 2, 2016.

[23]https://play.google.com/store. Accessed March 2, 2016.

http://www.springer.com/978-3-319-32965-9

The Conversational Interface
Talking to Smart Devices
McTear, M.; Callejas, Z.; Griol, D.
2016, XXII, 422 p. 76 illus., 29 illus. in color., Hardcover
ISBN: 978-3-319-32965-9