© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

7

Verbal Interaction

What is covered in this chapter:

- The complexities and challenges of human verbal interaction;
- The components of speech in human and human–robot interaction (HRI);
- The basic principles of speech recognition and application to HRI;
- Dialogue management systems in HRI;
- Speech generation in HRI, including the use of chat bots.

Imagine you come across a robot at your local electronics shop. It says "Hello" as you approach and asks you what you are looking for today. You rattle off "Oh, I don't know, maybe a camera for my daughter, some batteries, and just looking around, you know." As you await a response, there's an extended silence from the robot. Then it repeats its initial question, asking you to speak slower and from a closer distance. Is the robot broken? You approach another one of the store's robots, with similar results. Why are conversations with robots so frustrating? (This did, in fact, happen to one of the authors.)

Speech is perhaps the most obvious mode of communication among humans because it is both audible and explicit. It is also a common mode of communication designed into robots, both in terms of the speech produced by the robots and speech as input for robots. However, producing robot speech is much simpler than understanding human speech, which creates an imbalance between people's expectations and the robot's actual capabilities. In this chapter, we describe the main components of human speech and then discuss the mechanisms by which a robot can be prepared for verbal interaction.

7.1 Human–human verbal interaction

In human communication, speech serves various functions: it is used simply to convey information, but equally importantly, it also serves to create joint attention and a shared reality through communication. In addition to being an inherent part of our nature, speech is incredi-

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

Figure 7.1 The difficulties these two Peppers in a store in Tokyo had in communicating with passersby could have been due to the noisy environment or the diverse ways in which people communicate verbally.

bly complex and open to multiple interpretations. By a mere twist of intonation or shift in emphasis, the meaning of the same sentence can switch dramatically. For example, try to pronounce the following sentence eight times while putting emphasis on the next word each time, starting with the first word of the sentence, “she”:

She said she did not take his money.

By shifting the emphasis from one word to the next, what is inferred by the listener changes from a statement of belief (*She* said she did not take his money; apparently, someone else claimed otherwise) to disbelief (She *said* she didn’t take the money, but someone actually has seen her doing it), to an accusation (She said *she* didn’t [...], but someone else did), and so on.

Verbal communication is enriched by paralinguistic information as well, such as prosody and nonverbal behavior such as gaze, gestures, and facial expressions (see Chapter 6).

7.1.1 Components of speech

An *utterance* is the smallest unit in spoken language. Spoken language typically contains pauses between utterances, and an utterance is often less grammatically correct than a written sentence would be. This can become painfully clear when we read the transcript of a random

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

sentence from a conversation: whereas it takes no effort to understand what the person means when the person says it, the same sentence may appear incoherent when written down.

Spoken utterances can be short and consist of single words—such as *uhm*, *sure*, or *thanks*—or they can last for many minutes. Spoken language is often imperfect and has disfluencies, for example: “You know, I was, like, yeah, going to buy her, you know, something, but then I had, like, uhm, what the heck.”

Words are the smallest units that we can utter to convey meaning. In turn, *phonemes* are small units of sound that make up words, “pat,” for example, consists of three phonemes, [p] [a], and [t]. Changing a single one of them will change the meaning of the word; if the [p] is changed to a [b], we have a “bat.”

Conversational fillers make up part of speech without directly relating to a specific concept. They serve to keep a conversation going. For example, people utter “uh-huh” while listening to indicate that they are attending to and following the conversation. Conversational fillers are an important part of human verbal communication because they allow listeners to signal a broad range of responses (e.g., they are paying attention, they understand what the speaker means, they are surprised at a sudden twist in the story, or they share an emotion) without disrupting the flow of conversation. Such feedback increases the efficiency of verbal communication tremendously, and it enhances the experience of a shared reality between the speaker and the listener.

7.1.2 Written text versus spoken language

Written text and spoken utterances are vastly different. Whereas people expect rather strict adherence to grammatical rules and syntax in written text, they become much more liberal when talking. Because of the unidirectional nature of written communication, a written text needs to be prepared with a certain level of precision and refinement because it cannot be adjusted while it is being delivered.

Verbal communication, on the other hand, allows for many ways in which one can clarify any misunderstandings or obscurities while one is delivering the message. People usually quickly detect when the interaction partner does not understand the message in the intended way, and in response, they change their speech on the fly.

Natural and humanlike communication that runs smoothly is often crucial for human–robot interaction (HRI). However, in order to build natural-language interaction, many technical prerequisites have to be in place. These include the robot’s capability to transcribe speech into words, understand words by coming up with appropriate responses, and generate spoken language. The robot also often needs to be able to do this on the basis of verbal speech, which, as described previously, is more challenging than working with written text alone.

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

7.2 Speech recognition

Speech recognition is the recognition of spoken language by a computer and is also known as automated speech recognition (ASR) or speech-to-text (STT). Speech recognition is a process that takes a digital recording of speech and transcribes it. Speech recognition by itself does not understand or interpret what has been said. It merely converts a recorded fragment of speech into a written representation ready for further processing. Speech recognition has been mainly developed for controlling digital devices through spoken language or for dictation applications. Because of this, there is an assumption that the speech is recorded using a high-quality microphone, which is positioned close to the speaker in a relatively noise-free environment.

In HRI, these assumptions are often violated. When addressing a robot, the human conversation partner is often located at some distance from the robot, which has a negative impact on the quality of the recording. Signal processing and directional microphone arrays can alleviate this problem, but many robots are not equipped with such hardware. Due to the robot's microphone often not being located near the speaker's mouth, the microphone will also pick up sounds around the robot. Other people in the room talking, different sounds from the environment (e.g., a truck passing by outside, people walking around, or a cell phone ringing), and even mechanical noise from the robot itself all end up being recorded and provide a challenge for speech recognition. To avoid these problems, *close miking* is often used, where the user wears a lapel microphone or headset when talking to a robot.

The speech-recognition process requires a speech-recognition engine, software that has been trained to recognize one specific language. These are trained on thousands of hours of recorded and hand-transcribed speech, and they can handle only one language. Some speech-recognition engines are very specific and will only recognize brief commands or instructions specific to an application (e.g., recognizing spoken digits). Others are unconstrained and have been trained to recognize any possible spoken sentence. There are a few free, open-source speech-recognition engines, but the best-performing speech-recognition engines are commercial.

7.2.1 Basic principles of speech recognition

Speech recognition starts with a digital recording of speech, usually a recording of a single speaker. The recording is in the time domain: for every time step of the recording, for example, every 1/16,000th of a second, the sample contains the amplitude, or volume, of the recording. This is sufficient to replay the recording, but it is inconvenient for transcribing the speech into words. Thus, the recording is first converted to

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

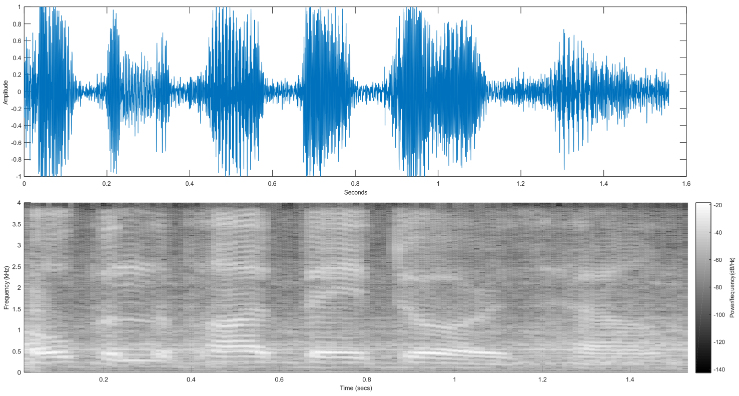

Figure 7.2 The speech sample "Open the pod bay doors, HAL" shown in the time and frequency domains. Speech recognition needs to transform these data into text.

the frequency domain. This means that it now shows how strongly certain frequencies are present in the signal at each time step. Phonemes look very different in the frequency domain—for example, an "o" has a different signature than an "a" in the frequency domain—and as such, they are easier to recognize with the use of an algorithm. Figure 7.2 shows a speech recording in both the time and frequency domains.

Until recently, speech-recognition engines used Gaussian mixture models and hidden Markov models to extract phonemes, words, and sentences from a speech recording. In essence, these approaches use probabilistic models of how phonemes and words can be strung together in words and sentences. The model knows that "robot" is a more likely transcription than "lobot" and that "the robot served the man" is more likely than "the robot swerved the nan."

In recent years, these probabilistic models have been replaced by deep neural networks (DNNs). These neural networks are similar in essence to the artificial neural networks that have been around since the 1960s, but their size is several magnitudes larger. A typical DNN can have hundreds of thousands of neurons and millions of connections between the neurons. Although these networks could not be trained in the past, new developments in algorithmic design and in computational hardware now allow the training of these networks to recognize spoken language relatively reliably. The performance of speech recognition using DNN has increased significantly compared to earlier methods. Not only has the rate of correctly recognized speech increased, but also speech-recognition engines can now increasingly deal with background noise, crowded environments, and ill-formed speech. They are also now speaker independent, meaning that the same speech-recognition model can deal with different speakers, including speakers of both genders.

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

7.2.2 Limitations

All speech-recognition engines still struggle with recognizing atypical speech. Speakers on which the models have been insufficiently trained, such as young speakers (Kennedy et al., 2017) or elderly speakers, still provide a challenge. Also, the local dialects of languages or nonnative speakers will often result in severely reduced recognition performance. The acoustic environment also is a determining factor. Noisy, reverberating, or crowded spaces will decrease ASR performance. Proper nouns, such as Margaret or Launceston Street, are also likely to be incorrectly picked up by speech recognition.

Constraining what needs to be recognized could increase the performance of the speech engine. To do so, most ASR engines allow the programmer to set constraints on what should be recognized, for example, digits from 0 to 10 or simple commands. Although constrained ASR can handle atypical speech with some success, the current state of the art still does not allow spoken interactions with target persons from different backgrounds.

7.2.3 Practice in HRI

Numerous speech-recognition engines are available. Speech recognition using DNNs is, due to the computational resources needed to store and compute through the networks, usually available as a remote service. These cloud-based solutions allow you to send a recorded speech fragment over the internet, and the transcribed speech is returned soon after. Next to the best and most up-to-date performance offered by cloud-based services, cloud-based recognition also frees up computational resources on the robot, allowing the robot to have a relatively low-cost computational core. If the nature of the application does not allow the use of cloud-based ASR, for example, because the robot does not have a reliable, always-on internet connection, there are on-board speech-recognition solutions that use a reduced DNN or first-generation approaches to speech recognition. Their performance is, however, lower than that of the cloud-based services.

Many big software companies provide cloud-based speech-recognition services. Google, IBM, Microsoft, and Nuance all offer pay-per-use cloud speech recognition. Recognizing a single speech sample is often free for low-frequency use, but costs are on the order of 1 cent per recognition event. There are a few free open-source alternatives, such as the Mozilla Foundation's Common Voice initiative, which builds an open and publicly available data set of voices to train speech-enabled applications, and its DeepSpeech recognition engine.

Speech-recognition engines generally have a simple-to-use application programming interface (API), allowing the programmer to quickly

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

integrate speech recognition on the robot. Next to the transcribed sentence, ASR engines will often also return a confidence value for the transcribed sentence, giving a measure of how confident the engine is about the recognized speech. Some engines will even return alternative transcriptions, again with confidence values.

7.2.4 Voice-activity detection

In some HRI applications, speech recognition is difficult due to the presence of noise, for example, because the robot is located in a crowded public space. Still, we can make a robot respond, albeit in a somewhat limited way, to people talking by using voice-activity detection.

Voice-activity detection (VAD) is often part of ASR, and it distinguishes speech from silence as well as other acoustic events. There is VAD software that can, for example, tell the difference between music playing and someone talking.

In HRI, VAD is used to give the user the impression that the robot is listening and can be used to implement spoken language turn-taking without actually recognizing or understanding the user's speech. In recent years, deep learning has also improved VAD performance. The free OpenSmile software package (Eyben et al., 2013) is currently leading in terms of performance. In combination with sound-source localization, whereby two or more microphones are used to pick up where a sound is coming from, we can even let the robot look at who is speaking.

7.2.5 Language understanding in HRI

A common misconception is that speech recognition also means that the speech is "understood" by the computer. It is not. Extracting semantic content from spoken language is particularly challenging, and a range of approaches exist that try to extract meaning from text, from broad semantic content to very specific content instructions.

Sentiment analysis, which matured as a way to analyze messages on social media, can be used to extract the affect contained in an utterance. Sentiment-analysis software often returns a scalar value denoting how negative or positive a message is. Although sentiment analysis is optimized for written language, in spoken language, we also have access to the way in which a message is delivered. Prosody and amplitude give us insight into the affect of the message: you do not need to speak the language to hear that the speaker is happy or agitated. Much in the same way, sentiment analysis and emotion from speech can roughly classify the affective state of the speaker.

More advanced methods, called natural-language understanding (NLU), will extract key words from the transcribed sentence. Elements such as

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

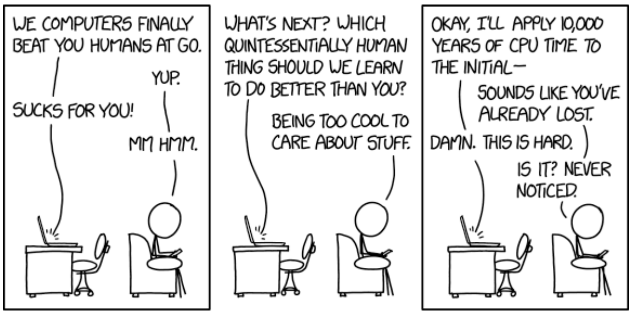
WE COMPUTERS FINALLY BEAT YOU HUMANS AT GO.
YUP.
SUCKS FOR YOU!
MM HMM.
WHAT'S NEXT? WHICH QUINTESSENTIALLY HUMAN THING SHOULD WE LEARN TO DO BETTER THAN YOU?
BEING TOO COOL TO CARE ABOUT STUFF.
OKAY, I'LL APPLY 10,000 YEARS OF CPU TIME TO THE INITIAL—
SOUNDS LIKE YOU'VE ALREADY LOST.
DAMN. THIS IS HARD.
IS IT? NEVER NOTICED.

Figure 7.3 It will likely take many years before artificial intelligence can successfully detect sarcasm. (Source: XKCD)

a command, location, person, event, or date can be picked out from a message, allowing the software to respond appropriately. These methods are often used in personal digital assistants. “Remind me to pick up the kids at 7 p.m.” is interpreted as a command to set a reminder at the time of 7 p.m. for the “pick up the kids” event. Those who frequently use digital assistants that extract information soon learn to conform to delivering information in ways that the computer can understand, and they settle on a manner of speaking that helps the computer perform satisfactorily.

Open-ended understanding of natural language is the biggest current challenge. Because conversational agents such as chat bots use conversational tricks to deal with ambiguous or poorly understood language, they might give the impression to the naive user that the agent understands what is being said. However, genuine understanding of natural language has not yet been achieved.

Even when people quickly realize that a robot does not understand figures of speech, they may still take some time to figure out how to phrase their message in such a way that the robot actually understands what they want. This is shown in the following dialogue that actually occurred between a human and a robot waiter, where the robot was programmed to take everything literally (Williams et al., 2018):

PARTICIPANT: Um, can I have one water?
ROBOT: Yes, that is permissible.
PARTICIPANT: Great.
ROBOT: Please tell me your order.
PARTICIPANT: Can I have one water?
ROBOT: Yes, that is permissible.
PARTICIPANT: Great, I'll take one water.
ROBOT: Thank you for sharing that interesting prediction. Please tell me your order.
PARTICIPANT: Can I order one water?
ROBOT: Yes, that is permissible.

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

PARTICIPANT: Okay, can you bring me one water?
ROBOT: Yes, I am able to do that.
PARTICIPANT: May you bring me one water?
ROBOT: Yes, I am able to do that.
PARTICIPANT: I would like to have one water.
ROBOT: Thank you for sharing that interesting fact. Please tell me your order.
[*16-second pause*]
PARTICIPANT: Can you bring me one water?
ROBOT: Yes, I am able to do that. Please tell me your order.
PARTICIPANT: May you please bring me one water?
ROBOT: Yes, I am able to do that.
PARTICIPANT: So can you do it?
ROBOT: Yes. Please tell me your order.
PARTICIPANT: Can you go inside and get the water for me?
ROBOT: Yes, I am able to do that. Please tell me your order.
PARTICIPANT: My order is water.
ROBOT: Okay.
[*Robot retrieves order for participant*]

7.3 Dialogue management

Dialogue management (DM) is the process that keeps a conversation, either written or spoken, flowing between a computer and a user. Key to dialogue management is that the interaction consists of several turns, instead of a single instruction. A command such as “play Bob Marley” does not require DM, whereas “I would like to order something” requires the agent to further explore what it is the user wants, which could result in a whole chain of to-and-fro questions on the exact nature of the order. The management of such a dialogue requires the dialogue manager to keep track of the *state* of the conversation. This state can be specific to the application or can be very generic. If the robot is handling a pizza order, then the state consists of the details of the pizza order, such as which size pizza the user wants or what toppings go on the pizza. The dialogue manager will try to complete the unknowns in the state from the user’s spoken utterances and will ask specific questions to fill in missing information before passing the order on.

7.3.1 Basic principle

Dialogue managers can range in complexity from systems that take the user through a very strict order of dialogue turns to systems that allow the user great liberty in the interaction. Restricted dialogue managers

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

are used for closed and well-contextualized tasks, such as registering guests, filling out forms, or taking orders. System-initiative dialogue managers leave no room for the user to change the course of the dialogue: the agent asks a number of questions to which the user responds. User-initiative dialogue managers allow the user to take the lead, and the system only intervenes when something is unclear or when more information is needed. Mixed-initiative systems offer a combination of both.

Simple dialogue managers will allow the programmer to design a dialogue script, which, in its simplest form, implements a linear script. Conditions and branches, which take the dialogue in a different direction, allow more flexibility. In essence, these dialogue managers are finite-state machines (FSMs), a set of instructions that define a flow and often refer to a computer program. More advanced systems can accommodate event-based DM, allowing the control flow to be interrupted by an event. Event-based dialogue managers can handle nonlinear dialogue flows, such as the user asking for the current time in the middle of placing a pizza order.

QiChat is the dialogue manager software used for Softbank Robotics robots, such as the Nao and Pepper robots. Here is an example of QiChat code. It listens for sentences by the user, such as greetings ("hi," "hello," and "hey there") and orders (e.g., "Do you have lemonade?").

```
topic:~introduction ()
language:enu
concept:(greetings) ^rand[hi hello "hey there"]
concept:(wine) [red white] wine
concept:(alcohol) [beer ~wine]
u:(~greetings) ~greetings
u:(do you have _~drink) yes, I have $1
u:(I want to drink something) do you want ~alcohol?
```

An example of the dialogue it produces is as follows:

ROBOT: Hey there.
USER: Hello.
ROBOT: Do you have white wine?
USER: Yes, I have white wine
ROBOT: I want to drink something.
USER: Do you want beer?
ROBOT: I want to drink something.
USER: Do you want red wine?

Still more advanced dialogue managers will use a planner that cap-

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

tures the state of the system and actions that can be taken to change the state. The advantage of using a planner is that the programmer is no longer required to write a dialogue script that captures every way in which the state needs to be filled; rather, the planner can search which actions are still needed to complete the state. So, rather than explicitly writing the questions the robot needs to ask to complete a pizza order, the planner knows that the state of a pizza contains variables such as size, toppings, and delivery time and will find the actions, in this case questions, needed to complete any missing information in the pizza order.

7.3.2 Practice in HRI

There are several commercial solutions for DM; for example, companies that provide speech-recognition services will often provide DM together with speech production. Dialogue managers can range from very simple script-based services, allowing the programmer to implement linear linguistic interactions, to complex and rich services with planners. The most popular dialogue managers are event based because these have sufficient flexibility for most language-based commercial interactions. Dialogue managers, however, are not at all suitable to implement free-flowing and open conversation. Free linguistic conversation requires a large range of dialogue rules, and the dialogue script soon becomes unwieldy.

Turn-taking in HRI

Spoken dialogue with a robot will invite the user to take a more natural stance toward interaction, and as such, it might be necessary to introduce a number of factors that are also present in human interaction. One of those is *back-channeling*—the responses given by the listener during a conversation to signal that he or she is still engaged, such as "uh-huh" or "really?". When your conversation partner is visible, there is often nonverbal back-channeling, such as a brief nod or a smile. In personal assistants, this often takes the form of a visual signal, such as a throbbing light, but on robots, these back-channeling signals can mimic human signals. The robot can use verbal back-channel signals, from the nonlexical "uh-huh" and "hmm" utterances to the phrasal and substantive utterances such as "yeah" and "tell me more." The robot could augment these with signals, such as blinking lights or a gentle hum, to show that it is listening and paying attention. One of the problems in using back-channeling on robots is when to use a back-channeling signal because the timing is dependent on speaker verbal and nonverbal cues. For example, Park et al. (2017) showed that a robot using a back-channel prediction model that provided contingent back-channel signals was preferred by children.

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

The role of timing

Timing is critical in natural interaction: when a response is delayed, this is seen as disturbing, whereas a very quick response is often seen as insincere (Sacks et al., 1974; Heldner and Edlund, 2010). The timing of the response also depends on other factors. Increased cognitive load slows the response; yes/no answers have a faster response time than responses that require a fully formed reply (Walczyk et al., 2003). An analysis of telephone conversations showed that "yes" answers to a question take on average just 100 ms, whereas responses to undesired offers take on average almost 500 ms (Strömbergsson et al., 2013). A response given before the end of a question shows how human conversational partners anticipate questions and utter a response before the question is finished.

Computers are significantly slower than people in issuing dialogue responses. Due to the sequential processing chain in DM, a robot often needs several seconds before a response is formulated. Silences can be filled with conversational fillers or visual signals, signaling to the user that the robot is formulating a response. However, these are poor substitutes for prompt turn-taking, and considerable effort is being put into reducing the response delay in natural-language interaction. Just-in-time speech synthesis, where the robot starts speaking before having a plan of how to finish the sentence, seems promising, as does incremental spoken-dialogue processing, which works along the same principle as already-taken actions in response to spoken instructions before the instructions have been finished (Baumann and Schlangen, 2012).

7.4 Speech production

The final step in natural language interaction is converting a written response of the system into speech. For this, we need speech production, also known as speech synthesis or text-to-speech (TTS).

Speech production has seen impressive progress in recent years. In the 1990s, only voices that sounded tinny were available. Now, nearly 30 years later, we have artificial speech production that is almost indistinguishable from human speech. The two established methods for generating artificial speech are concatenative and parametric TTS. In concatenative speech production, an actor's voice has been recorded and cut into phonemes, and these are then "glued" together and smoothed at the seams to provide natural-sounding speech (Hunt and Black, 1996). In parametric TTS, a model is trained to produce acoustic speech parameters from text (Zen et al., 2009). Although concatenative models sound natural, they have little flexibility, and new voices require completely new recording and training of the TTS model. Parametric TTS is more flexible and allows for customization of the voice and prosody, at

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

the expense of naturalness. Recent advances have overcome these limitations by training generative deep neural networks (DNNs). Van den Oord et al. present a DNN model that produces speech that is virtually indistinguishable from human speech and even includes breathing and lip-smacking (van den Oord et al., 2016). This model has been adopted by Google as the voice of its digital assistant.

7.4.1 Practice in HRI

A wide selection of speech-production software is currently available, from free solutions to bespoke commercial software with voices customized to specific applications.

TTS engines

The simplest TTS engines have a small computational footprint and can run on cheap robot hardware. The most natural-sounding TTS engines use DNNs and are cloud based. Depending on the application, some TTS engines not only convert text into a speech file but also provide timing information for phonemes, which can be used to animate a robot. It might be necessary for the speech to be synchronized with facial animations or gestures on the robot, and timing information will allow for precise synchronization between the speech and the animations.

In HRI, it is important to consider which voice fits the robot and its application. A small robot requires a voice that matches its appearance, rather than a commanding baritone. In some cases, though, it might be important to match the sound of the voice to the fact that it emanates from a robot: a natural-sounding TTS engine might sit uneasily on an artificial agent. At the same time, research by Eyssel et al. (2012a) has shown that the type of voice affects the social perception of social robots. For example, robots with a male voice are anthropomorphized and evaluated more favorably by men than by women, and vice versa.

Some limitations to speech production still exist. Adaptive prosody and emotion, although actively being researched, are not commonly available on TTS engines. Also, synthesized voices do not adapt to the context in which they are being used. When the room is quiet, there is little need for the robot to have a booming voice, whereas a robot addressing a crowd at an exhibition would do well to adapt its rate of speech and volume to increase its intelligibility.

Chat bots

Chat bots are computer programs intended to converse with the user, typically by using written text. These systems are often implemented as a web application in which users enter text on a web page and the server responds to every text entry. These chat bots often have a specific

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

goal, such as to provide technical support or answer questions about the products of a company. Chat bots can become full sales agents or customer support agents. These agents are normally constrained regarding the topics to which they can respond in a meaningful way. More recently, chat bots have become speech-enabled. Chat bots such as Siri (Apple), Cortana (Microsoft), Alexa (Amazon), and Bixby (Samsung) now respond to simple spoken commands and respond with spoken text.

A second type of chat bot is general-purpose agents that try to respond to any utterance. They achieve this, on the one hand, by having thousands of hand-crafted rules on how to respond to often-occurring utterances and, on the other hand, by maintaining a database of all previous conversations, often learning from how users responded in the past to a given utterance. The ultimate goal is to create a chat bot that is indistinguishable from a human—users would no longer be able to tell if they are talking to a computer or a human. Controlled tests are being set up in an annual competition, and the most convincing chat bot receives the Loebner Prize. This test is often called a Turing test, named after the famous computer scientist Alan Turing, who proposed such a test as a measure for the intelligence of a computer (Turing, 1950).

The list of chat bots created by major information technology (IT) companies, such as Apple, Microsoft, Google, Amazon, and Facebook, indicates that there is already considerable interest in natural-language technology, and many companies make their technology available for developers. Google is offering its Cloud Speech application programming interface (API), Microsoft is pitching its Cognitive Services, and Amazon offers its Alexa set of tools to build voice-based services.

The availability of these services means it is no longer necessary to create your own software for speech recognition, understanding, or synthesis. Instead, developers can use online services for their robots. The audio signal recorded through the robot's microphone is streamed in real-time to the company's servers, and they send back the recognized text while the user is still talking. Similarly, these services can be used not only to recognize the spoken text but also to respond to the meaning of the text. The systems can, for example, identify entities, syntax, sentiments, and categories. This all helps the robot to better respond to the utterances of the users. These companies also offer speech-synthesis tools. The robot sends what it wants to say to a server and receives back the audio signal that the robot then plays on its loudspeakers.

It is much harder for a human to learn a new language than for a computer to do so. Still, artificial languages, such as Esperanto, have been developed to overcome some of the inherent problems

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

with learning natural languages. These constructed languages serve different purposes:

- Engineered languages—experimentation in logic, philosophy, or linguistics (Loglan, ROILA)
- Auxiliary languages—developed to help in the translations between natural languages (Esperanto)
- Artistic languages—created to enrich fictional worlds (Klingon, Elfish, or Dothraki)

The RObot Interaction LAnguage (ROILA) was developed for HRI, in particular to facilitate the problems that speech-recognition accuracy encounters (Stedeman et al., 2011). The words of this language have been designed to sound most distinct from each other, making it much easier for automated speech recognition to correctly identify the spoken words. “Go forward” in ROILA is “kanek koloke”; “go back” is “kanek nole.”

7.5 Conclusion

Despite being the most obvious form of communication among humans, language is very complex, not only due to the large number of words people use daily but also because their meaning and significance change based on various contextual factors (e.g., relationships between speakers, task, prosody). Creating robots that can engage in this rich and diverse form of communication is a necessary goal for HRI, and technical tools available for speech analysis, synthesis, and production enable some degree of verbal HRI that does not need to be developed from scratch. Open-ended, natural-language conversation is still not possible, but verbal interaction in more constrained contexts can be successfully applied on robotic platforms.

Questions for you to think about:

- Imagine a social robot that needs to perceive all of the utterances you speak at your home every day, and think of a list of words (dictionary) for ASR. How long would this list need to be for the robot to be able to understand your everyday conversations?
- Consider the difference in how you say “yes” willingly versus reluctantly. How would you make a robot respond appropriately to such different modes of speaking?
- What are some problems that can emerge in relation to the important role of timing in human–robot interactions? How are these solved in other social interactions where the interactants

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.

© copyright by Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic 2019.
https://www.human-robot-interaction.org

miss out on social cues (e.g., in texting, or when there is time delay on Skype calls)?

Future reading:

- Amir Aly and Adriana Tapus. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, pages 325–332, Piscataway, NJ, USA, 2013. IEEE Press. ISBN 978-1-4673-3055-8. doi: 10.1109/HRI.2013.6483606. URL `https://doi.org/10.1109/HRI.2013.6483606`
- J. Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. *Embodied conversational agents*. MIT Press, Cambridge, MA, 2000. ISBN 9780262032780. URL `http://www.worldcat.org/oclc/440727862`
- Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura de Ruiter. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *Robot and human interactive communication (RO-MAN)*, pages 851–857. IEEE, 2012b. doi: 10.1109/ROMAN.2012.6343858. URL `https://doi.org/10.1109/ROMAN.2012.6343858`
- Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. A communication robot in a shopping mall. *IEEE Transactions on Robotics*, 26(5):897–913, 2010. doi: 10.1109/TRO.2010.2062550. URL `https://doi.org/10.1109/TRO.2010.2062550`
- Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015. ISSN 0921-8890. doi: 10.1016/j.robot.2014.09.031. URL `https://doi.org/10.1016/j.robot.2014.09.031`
- Michael L. Walters, Dag Sverre Syrdal, Kheng Lee Koay, Kerstin Dautenhahn, and R. Te Boekhorst. Human approach distances to a mechanical-looking robot with different robot voice styles. In *Robot and human interactive communication (RO-MAN)*, pages 707–712. IEEE, 2008. doi: 10.1109/ROMAN.2008.4600750. URL `https://doi.org/10.1109/ROMAN.2008.4600750`

This material has been published by Cambridge University Press as Human Robot Interaction by
Christoph Bartneck, Tony Belpaeime, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Sabanovic.
ISBN: 9781108735407 (http://www.cambridge.org/9781108735407).
This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works.