

Real-Time Coordination in Human-Robot Interaction Using Face and Voice

Gabriel Skantze

■ *When humans interact and collaborate with each other, they coordinate their turn-taking behaviors using verbal and nonverbal signals, expressed in the face and voice. If robots of the future are supposed to engage in social interaction with humans, it is essential that they can generate and understand these behaviors. In this article, I give an overview of several studies that show how humans in interaction with a humanlike robot make use of the same coordination signals typically found in studies on human-human interaction, and that it is possible to automatically detect and combine these cues to facilitate real-time coordination. The studies also show that humans react naturally to such signals when used by a robot, without being given any special instructions. They follow the gaze of the robot to disambiguate referring expressions, they conform when the robot selects the next speaker using gaze, and they respond naturally to subtle cues, such as gaze aversion, breathing, facial gestures, and hesitation sounds.*

For a long time, science fictions writers and scientists have been entertaining the idea of the speaking machine — an automaton, computer, or robot that you could interact with by means of natural language, just like we communicate with each other. In his seminal paper *Computer Machinery and Intelligence*, Alan Turing argued that this ability would indeed be a defining feature of intelligence (Turing 1950). If a human subject would sit at a terminal and chat with an unknown partner without being able to tell whether it is another human or a machine, we would have managed to create artificial intelligence. Since then, this thought experiment has been followed up by attempts at actually building such a system, from the artificial psychotherapist Eliza (Weizenbaum 1966), to customer service chatbots on websites, and now (with the addition of speech) voice assistants in our mobile phones, such as Apple's Siri and Microsoft's Cortana. While this development has indeed shown impressive progress in terms of user acceptance (perhaps mostly thanks to breakthroughs in speech recognition), these systems rely on a fairly simplistic model of human interaction, where two interlocutors exchange utterances

using a very strict turn-taking protocol. In a written chat, the end of a turn is typically marked with the return key, and voice assistants typically use a button or a key word (like Amazon's "Alexa") to initiate a turn, and then a long pause to mark the end.

Contrary to this, most conversational settings in everyday human interaction do not have such strict protocols, with the exception of very special situations such as communication over a walkie-talkie. Spoken interaction is typically coordinated on a much finer level, and humans are very good at switching turns with very short gaps (around 200 ms) and little overlap. Humans also give precisely timed feedback in the middle of the interlocutor's speech in the form of very short utterances (so-called backchannels, such as "mhm") or head nods. Another notable property of everyday human interaction is that it is often physically situated, which means that the space in which the interaction takes place is of importance. In such settings, there might be several interlocutors involved (so-called multiparty interaction), and there might be objects in the shared space that can be referred to. Also, the interaction might revolve around some joint activity (such as solving a problem), and the speech has to be coordinated with this activity. An important future application area for spoken language technology where all these issues will become highly important is human-robot interaction. Robots of the future are envisioned to help people perform tasks, not only as mere tools, but also as autonomous agents interacting and solving problems together with humans.

Another notable limitation with chat bots and voice assistants of today is that they almost exclusively focus on the verbal aspect of communication, that is, the words that are written or spoken. But human communication is also filled with nonverbal signals. It is important not just which words are spoken, but also how they are spoken — something speech scientists refer to as prosody (the melody, loudness, and rhythm of speech). Depending on the prosody, the speaker can be perceived as certain or uncertain, and utterances can be perceived as statements or questions. There are also other nonverbal aspects of speech that have communicative functions, such as breathing and laughter. Another aspect that is typically missing is the face, which includes important signals such as gaze, facial expressions, and head nods. What is especially interesting with these nonverbal signals, which will be the focus of this article, is that they are highly important for real-time coordination. Thus, if a robot is supposed to be involved in more advanced joint activities with humans, it should be able to both understand and generate nonverbal signals.

However, just because we manage to implement these things in social robots, it is not certain that humans will display these behaviors toward the robot, and react to the robot's nonverbal behavior in

an expected way. Also, processing these signals and making use of them in a spoken dialogue system in real time is a nontrivial task. In this article, I will summarize some of the results from several studies done at KTH to address these questions.

Research Platform

Before discussing the challenges of real-time coordination in human-robot interaction, I will present the research platforms that we have developed at KTH: the robot head *Furhat* and the interaction framework *IrisTK*. I will also present two different application scenarios that we have developed, which pose different types of challenges when it comes to modeling turn-taking, feedback, and joint attention in human-robot interaction.

The *Furhat* Robot Head

The face carries a lot of information — it provides the speaker with a clear identity, the lip movements help the listener to comprehend speech, facial expressions can signal attitude and modify the meaning of what we say, head nods can provide feedback, and the gaze helps the listener to infer the speaker's visual focus of attention. Until recently, the standard solution for giving conversational agents a face has been to use an animated character on a display, so-called embodied conversational agents (or ECAs for short). The importance of facial and bodily gestures in ECAs has been demonstrated in several studies (Cassell et al. 2000). However, when it comes to physically situated interaction, animated characters on two-dimensional (2D) displays suffer from the so-called Mona Lisa effect (Al Moubayed, Edlund, and Beskow 2012). This means that it is impossible for the observer to determine where in the observer's physical space the agent is looking. Either everyone in the room will perceive the agent as looking at them, or nobody will, which makes it impossible to achieve exclusive mutual gaze with just one observer. This has important implications for many human-robot interaction scenarios, where there may be several persons interacting with the robot, and where the robot may look at objects in the shared space.

To combine the advantages of animated faces with the situatedness of physical robotic heads, we have developed a robot head called *Furhat* at KTH (Al Moubayed, Skantze, and Beskow 2013), as seen in figures 1–3. An animated face is back-projected on a static mask, which is in turn mounted on a mechanic neck. This allows *Furhat* to direct his gaze using both head pose (mechanic) and eye movements (animated). Compared to completely mechatronic robot heads, this solution is more flexible (the face can easily be changed by switching mask and animation model), and allows for very detailed facial expressions without generating noise. To validate that this solution does not suffer from the Mona Lisa effect,

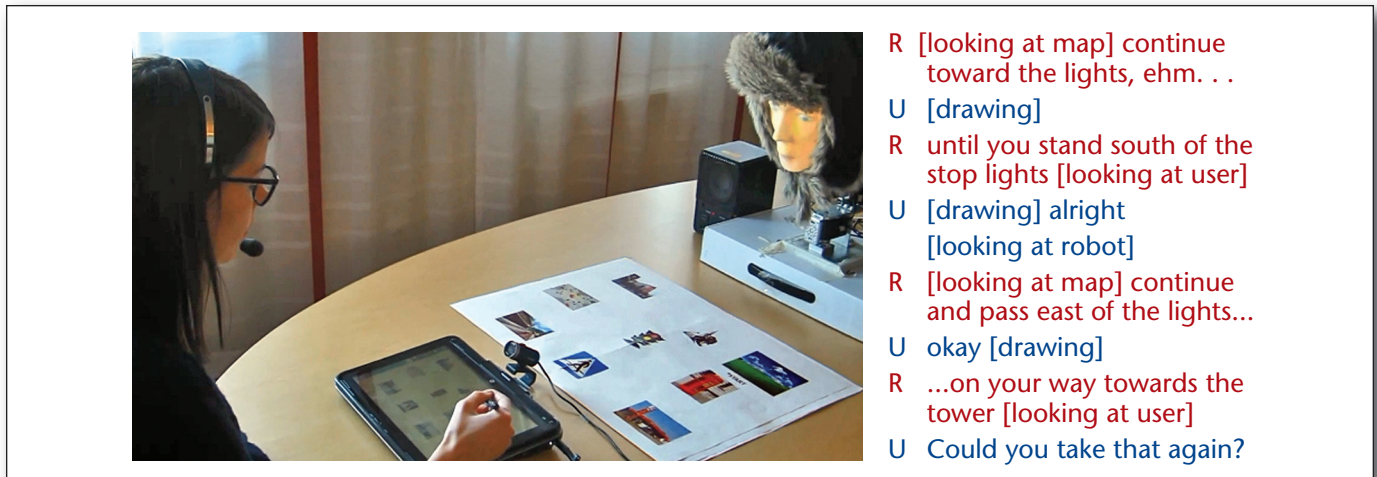


Figure 1. Furhat Instructing a Human Subject on How to Draw a Route on a Map.

we have done a series of experiments, where we systematically compared Furhat with an animated agent on a 2D display, and found that Furhat can indeed achieve mutual gaze in multiparty interaction, and that subjects can determine the target of Furhat's gaze in the room nearly as well as the gaze of a human. Furthermore, we have shown that Furhat's animated lip movements improve speech comprehension significantly under noisy conditions (Al Moubayed, Skantze, and Beskow 2013).

Interaction Scenarios

In this article, I will discuss results from two different human-robot interaction scenarios. In the first scenario, depicted in figure 1, Furhat instructs a human on how to draw a route on a map (Skantze, Hjalmarsson, and Oertel 2014). A human subject and the robot are placed face to face with a large printed map on the table between them, which constitutes a target for joint attention. The robot describes the route, using the landmarks on the map, and the subject is given the task of drawing the route on a digital map in front of her. In this task, the robot has to coordinate the information delivery with the human's execution of the task (drawing the route). To this end, the robot has to "package" the instructions in appropriately sized chunks and invite feedback from the user (Clark and Krych 2004). The user then has to follow these instructions and give feedback about the task progression. Together, they continuously have to make sure that they attend to the same part of the map. The system was tested with 24 recruited participants.

In the second scenario, depicted in figure 2, two humans play a collaborative card-sorting game together with Furhat (Skantze et al. 2015). The task could for example be to sort a set of inventions in the order they were invented, or a set of animals based on how fast they can run. Since the game is collabora-

tive, the humans have to discuss the solution together with each other and Furhat. However, Furhat is programmed not to have perfect knowledge about the solution. Instead, Furhat's behavior is motivated by a randomized belief model. This means that the humans have to determine whether they should trust Furhat's belief or not, just like they have to do with each other. Similar to the first scenario, the touch table with the cards constitutes a target for joint attention. However, they are different in that this task requires coordination among three participants (so-called multiparty interaction), and is of a more open, conversational nature, where the participants' roles are more symmetrical. This system was exhibited during one week at the Swedish National Museum of Science and Technology in November 2014, where we recorded almost 400 interactions with users from the general public, including both children and adults.¹

Modeling the Interaction Using IrisTK

For a robot to engage fully in face-to-face interaction, the underlying system must be able to perceive, interpret, and combine a number of different auditory and visual signals and be able to display these signals in the robot's voice and face. To facilitate the implementation of such systems, we have developed an open source framework called IrisTK,² which provides a modular architecture and a set of modules for modeling human-robot interaction (Skantze and Al Moubayed 2012). It has been used to implement a number of different systems and experimental setups, including the two settings described above. I will only give a brief overview here, but the interested reader can refer to Skantze, Johansson, and Beskow (2015) for a more detailed description of how it was used in the card-sorting game.

The most important components are schematically illustrated in figure 3. The speech from the two



U-1 I wonder which one is the fastest [looking at table]
 U-2 I think this one is fastest, what do you think? [looking at robot]
 R I'm not sure about this, but I think the lion is the fastest
 U-1 Okay [moving the lion]
 R Now it looks better
 U-2 Yeah... How about the zebra?
 R I think the zebra is slower than the horse. What do you think? [looking at U-2]
 U-2 I agree

Figure 2. Two Children Playing a Card-Sorting Game with Furhat.

U-1 and U-2 denote the two users.

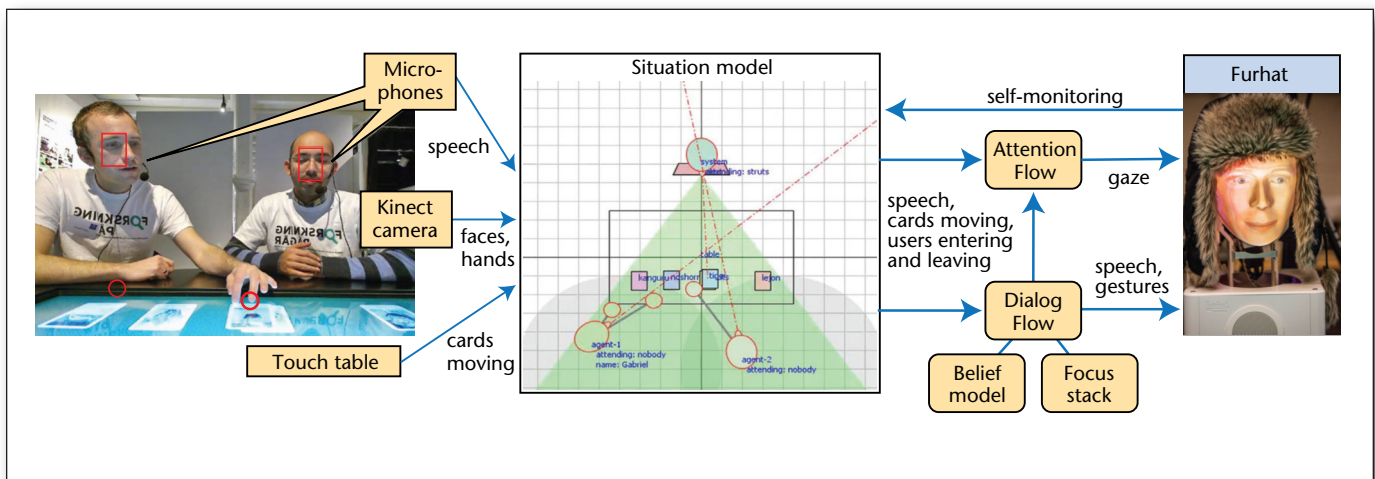


Figure 3. Overview of the Different Components and Some of the Events Flowing in the System.

users is picked up either by close talking microphones or by a microphone array and is recognized and analyzed in parallel, which allows Furhat to understand both users, even when they are talking simultaneously. To visually track the users that are in front of Furhat, a Microsoft Kinect camera is used, which provides the system with information about the position and rotation of the users' heads (as a rough estimation for their visual focus of attention). These inputs, along with the movement of the cards on the touch screen table, are sent to a situation model, which merges the multimodal input and maintains a three-dimensional (3D) representation of the situation. A dialogue flow module orchestrates

the spoken interaction, based on events from the situation model, such as someone speaking, shifting attention, entering or leaving the interaction, or moving cards on the table. An attention flow module keeps Furhat's attention to a specified target (a user or a card), by consulting the situation model.

Coordination Mechanisms in Spoken Interaction

Many human social activities require some kind of turn-taking protocol, that is, to negotiate the order in which the different actions are supposed to take place, and who is supposed to take which step when.

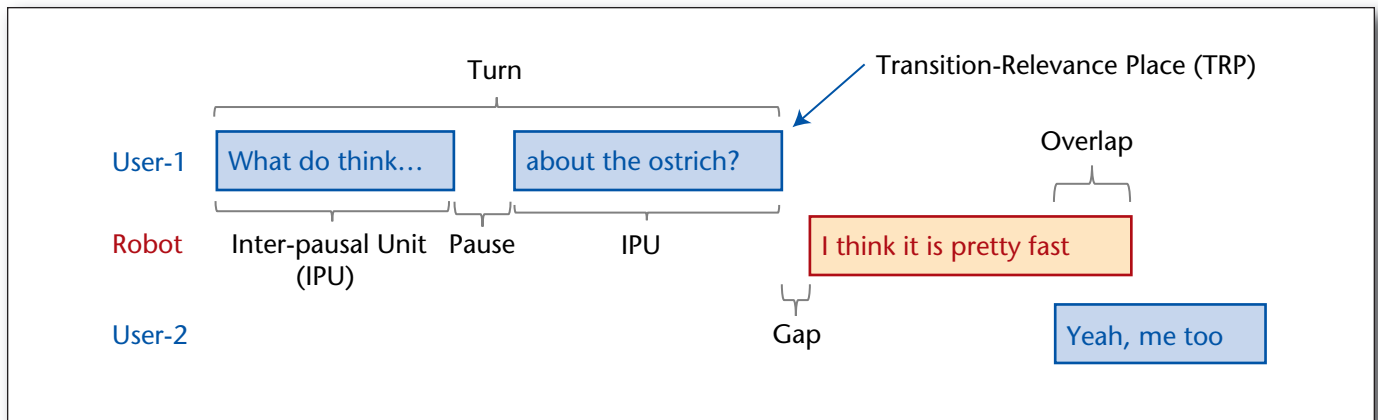


Figure 4. Important Concepts When Modeling Turn-Taking.

This is obvious when for example playing a game or jointly assembling a piece of furniture, but it also applies to spoken interaction. Since it is difficult to speak and listen at the same time, speakers in dialogue have to somehow coordinate who is currently speaking and who is listening. Studies on human-human interaction have shown that humans coordinate their turn-taking and joint activities using a number of sophisticated coordination signals (Clark 1996).

Some important concepts in this process are shown in figure 4, which illustrates a possible interaction from the card-sorting game described above. From a computational perspective, a useful term is *inter-pausal unit (IPU)*, which is a stretch of audio from one speaker without any silence exceeding a certain amount (such as 200 ms). These can relatively easily be identified using voice activity detection. A turn is then defined as a sequence of IPUs from a speaker, which is not interrupted by IPUs from another speaker. At certain points in the speech, there are transition relevance places (TRPs), where a shift in turn could potentially take place (Sacks, Schegloff, and Jefferson 1974). As can be seen, there might be pauses within a turn, where no turn-shift is intended, but there might also be overlaps between IPUs and turns. Even if gaps and overlaps are common in human-human interaction (Heldner and Edlund 2010), humans are typically very good at keeping them short (often with just a 200 ms gap).

Traditionally, spoken dialogue systems have rested on a very simplistic model of turn-taking, where a certain amount of silence (say 700–1000 ms) is used as an indicator for transition-relevance places. The problem with this model is that turn-shifts often are supposed to be much more rapid than this, and that pauses within a turn often might be longer. This means that the system will sometimes appear to give sluggish responses, and sometimes interrupt the user. Thus, silence is not a very good indicator for a turn-shift. Another solution would be to make a continu-

ous decision on when to take the turn (say every 100 ms), or break up the user's speech into several IPUs using much shorter pause thresholds (such as 200 ms), and then try to identify whether the user is yielding or holding the turn after each IPU. But what should this decision be based on?

Several studies have found that speakers use their voice and face to give turn-holding and turn-yielding cues (Duncan 1972; Koiso et al. 1998; Gravano and Hirschberg 2011). For example, an IPU ending with an incomplete syntactic clause ("how about...") or a filled pause ("uhm...") typically indicates that the speaker is not yielding the turn. But as the example in figure 4 illustrates, it is not always clear whether syntactically complete phrases like "what do you think" are turn-final or not. Thus, speakers also use prosody (that is, how the speech is realized) to signal turn-completion. Three important components of prosody are pitch (fundamental frequency), duration (length of the phonemes) and energy (loudness). A rising or falling pitch at the end of the IPU tend to be turn-yielding, whereas a flat pitch tends to be turn-holding. The intensity of the voice tends to be lower when yielding the turn, and the duration of the last phoneme tends to be shorter. By breathing in, the speaker may also signal that she is about to speak (thus holding the turn) (Ishii et al. 2014). Gaze has also been found to be an important cue — speakers tend to look away from the addressee during longer utterances, but then look back at the addressee toward the end to yield the turn (Kendon 1967). Gestures can also be used as an indicator, where a non-terminated gesture may signal that the turn is not finished yet. A summary of these cues is presented in table 1. Another important aspect to take into account is the dialogue context. If a fragmentary utterance (like "the lion") can be interpreted as an answer to a preceding question ("which animal do you think is fastest?"), it is probably turn-yielding, but might otherwise just be the start of a longer utterance.

	Turn-yielding cue	Turn-holding cue
Syntax	Complete	Incomplete, Filled pause
Prosody — Pitch	Rising or Falling	Flat
Prosody — Intensity	Lower	Higher
Prosody — Duration	Shorter	Longer
Breathing	Breathe out	Breathe in
Gaze	Looking at addressee	Looking away
Gesture	Terminated	Nonterminated

Table 1. Turn-Yielding and Turn-Holding Cues Typically Found in the Literature.

Detecting Coordination Signals

It is important to note that the cues listed in Table 1 are very schematic — all these cues do not conform to these principles all the time. However, studies on human-human dialogue have shown that the more turn-yielding cues are presented together, the more likely it is that the other speaker will take the turn (Duncan 1972; Koiso et al. 1998; Gravano and Hirschberg 2011). In this section, I will discuss how machine learning can be used to combine and classify the rich source of multimodal features picked up by the sensors in IrisTK, allowing the robot to coordinate the interaction with humans.

Knowing When to Speak in Multiparty Interaction

In a multiparty setting such as the card-sorting game, the system does not only have to determine whether the user is yielding the turn or not, but also to whom the turn is yielded. If it is yielded to the other human, the robot should not take the turn. To do this, it is important to be able to detect the addressee of user utterances. Other researchers have found that this can be done by combining several different multimodal cues, using machine learning (Katzenmaier et al. 2004; Vinyals, Bohus, and Caruana 2012). However, these studies have mostly been done in interaction scenarios where the robot has a very clear role, such as a butler or a quiz host. In such settings, the user is typically either clearly addressing the robot or another human. In the card-sorting game scenario, where the robot is involved in a collaborative discussion, it is often much harder to make a clear binary decision, both regarding whether the turn was yielded or not, and whether a particular speaker was being addressed (Johansson and Skantze 2015). We therefore chose to combine these two decisions into one: Should the robot take the turn or not? If not, it is either because the current speaker did not yield the turn, or because the turn was yielded to the other human. There are also clear cases where Furhat is

“obliged” to take the turn, for example, if a user looks at Furhat and asks a direct question. In between these, there are cases where it is possible to take the turn “if needed,” and cases where it is appropriate to take the turn, but not obligatory. To create a gold standard for these decisions, we gave an annotator the task of watching videos of the interactions from Furhat’s perspective, and choose the right turn-taking decision after each IPU, using a scale from 0 to 1 (where 0 means “don’t take the turn” and 1 means “obliged to take the turn”). The result of this annotation (the histogram for 10 dialogues) is shown in figure 5. To see if we could build a model for predicting this decision using multimodal features, we first trained an artificial neural network to make a decision between the two extreme categories: “Don’t” and “Obliged” (Johansson and Skantze 2015). As can be seen from the results in figure 5, head pose (as a proxy for gaze) is a fairly good indicator, which might not be surprising, since gaze can both serve the role as a turn-yielding signal and as a device to select the next speaker. But it also shows that combining features from different modalities improves the performance significantly, in line with studies on human-human interaction. Another observation is that many of the features seem to be redundant. It is also interesting that card movement is a useful feature — if the user was not done with the current movement, the turn was not typically yielded, which is similar to how gestures can be informative (see table 1). To complement this binary classifier, we also built a regression model (using Gaussian processes) to predict the continues outcome on the whole turn-taking spectrum, which yielded an R-value of 0.677, when all features were combined.

In the end, the system will have to make a binary decision of whether to take the turn or not, and so far we have only used the binary classifier for making this decision. The decision should, however, ultimately also take into account what the robot actually has to contribute, and how important this contribution is, not just to what extent the last turn was

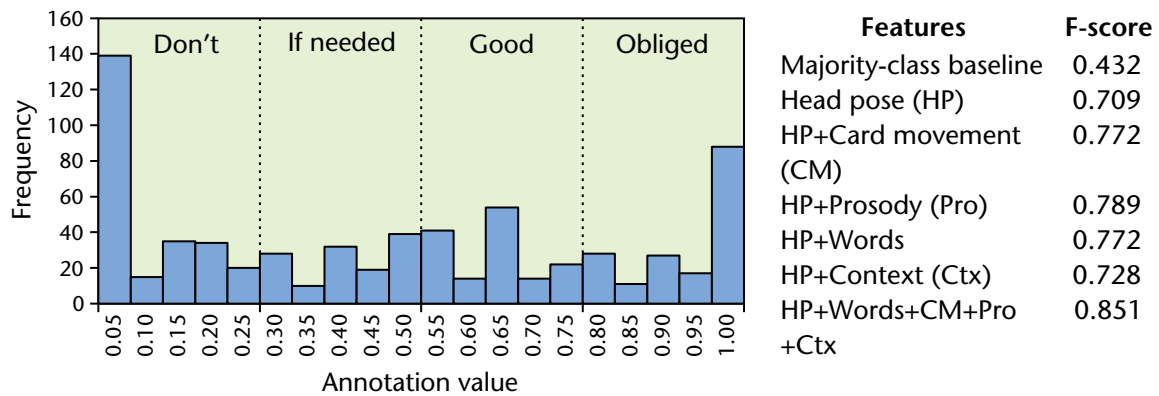


Figure 5. Annotation Result — The Histogram for 10 Dialogs.

Left: Histogram of annotated turn-taking decisions on a scale from 0 (must not take turn) to 1 (must take turn). Right: Prediction of Don't versus Obligated using an artificial neural network with different sets of features.

yielded or not. For future work, we therefore want to combine this utility with the outcome of the regression model, in a decision-theoretic framework. If the robot would have something very important to say, it might not matter whether it is a good place to take the turn or not. And the other way around, even if the robot does not have anything important to contribute, it might have to say something anyway, if it has an obligation to respond. Intuitively, this is the kind of decisions we as humans also continuously make when engaged in dialogue.

Recognizing Feedback from the User

As another example of how the system can detect coordination signals from the user, we will now turn to the map-drawing task described earlier (Skantze, Hjalmarsson, and Oertel 2014). In this scenario, the robot mostly has the initiative and is supposed to give route instructions in appropriately sized chunks, awaiting feedback from the user before it can continue. If we look at the user's verbal behavior, it mostly consists of very short feedback utterances, including "okay," "yes," "yeah," "mm," "mhm," "ah," "all right," and "oh." At a first look, it might seem like all these are just variations of the same thing. However, a more detailed analysis of the 1568 feedback utterances in the data revealed that these utterances do not always have the same meaning, and that the choice of verbal token and its prosodic realization was not arbitrary. Thus, the form of the feedback is somehow related to its function. One important aspect concerns the timing of the feedback in relationship with the drawing activity, which is illustrated in figure 6. A short feedback token such as "okay" might in fact mean "okay, I will do that," "okay, I

have done that now," "okay, I am doing that now," or "okay, I have already done that (in the previous step)." This distinction is important when timing the next piece of instruction from the robot. By relating the timing of the feedback with the timing of the drawing activity, we can automatically derive these functions and see how they relate to the form of the feedback. For example, a short, high intensity "yes" typically means, "I have already done that" (no need to draw anything), whereas a long "okaay" or "mm" with a rising pitch typically means, "I am doing that." As can be seen in the figure, the likelihood that the user will look up at the robot while giving this feedback is also different. When no more drawing is expected (the user wants the next piece of information), we can see that it is more common to look at the robot, thus in effect yielding the turn. The prosodic features to some extent also follow the turn-taking patterns listed in table 1, although the relationship is not so clear cut. To see whether a system could automatically detect and make use of these cues in the system, we built a logistic regression classifier that could predict the meaning of the feedback token with an F-score of 0.63 (which could be compared to a majority class baseline of 0.153).

These results show that the forms and functions of feedback are closely linked. There are of course many ways in which the functions of feedback can be categorized, where timing is one important aspect. Another aspect is the user's level of certainty, which we also found to be reflected by the choice of token, prosodic realization and gaze direction (Skantze, Hjalmarsson, and Oertel 2014). Feedback reflecting uncertainty is more often expressed with "ah" and "mm," and typically has a low intensity, longer dura-

Robot instruction	User response (FB = feedback)	Paraphrase meaning	Most common feedback tokens	Feedback prosody	Gaze at robot
Continue to the church	FB drawing	"I will do that"	okay, yes, mhm	Low intensity, Long duration, Flat pitch	34%
Pass east of the lights	FB	"I have already done that"	yes, okay, yeah	High intensity, Short duration, Flat pitch	55%
Go around the house	drawing FB	"I have done that now"	okay, yes, yeah	Medium intensity, Medium duration, Rising pitch	66%
Continue to the tower	FB drawing	"I am doing that now"	okay, yes, mm	Medium intensity, Long duration, Rising pitch	37%

Figure 6. How Prosody and Gaze in User Feedback Relates to the Coordination of the Ongoing Activity.

Drawing the route.

tion, and flat pitch. A system that can detect these functions in the user's feedback can better pace its instructions, and know when to elaborate on them further.

Generating Coordination Signals

So far, we have looked at examples of how the robot can perceive and interpret multimodal coordination signals from the user(s). But another important question is of course how the robot should be able to generate these signals using its voice and face. By generating the right coordination signals, the robot can both facilitate the interaction and make it more pleasant and less confusing for the user, but it can also be used to shape the interaction according to some criterion.

Guiding Joint Attention

As discussed, we have found in perception experiments that users can accurately determine the target of Furhat's gaze. This is important, since it potentially allows for joint attention between the user and the robot. However, it is not obvious whether humans will actually utilize the robot's gaze to identify referents in an ongoing dialogue, in the same way they do with other humans. In the map-drawing task, we investigated this by deliberately placing ambiguous landmarks (such as two different towers) on the map (Skantze, Hjalmarsson, and Oertel 2014). We then

experimented with three different conditions. First, a condition where Furhat was looking at the landmark he was referring to and looked up at the user at the end of each instruction (CONSISTENT). Second, a condition where Furhat randomly switched between looking in the middle of the map and looking up at the user (RANDOM). Third, a condition where we placed a cardboard in front of Furhat, so that the user could not see him (NOFACE). Since the users were drawing the route on a digital map, we could precisely measure the drawing activity (pixels per second) during the course of the instructions. The average drawing activity during ambiguous instructions is illustrated in figure 7. The CONSISTENT gaze clearly helped the user to find the object that was being referred to, which is indicated by the increased drawing activity during the pause. It is interesting to note that the RANDOM condition was in fact worse than the NOFACE condition, probably because the user spent time trying to utilize the robot's gaze (which didn't provide any help in that condition). This shows that humans indeed try to make use of the robot's gaze, and can benefit from it, if the gaze signal is synchronized with the speech in a meaningful way.

Selecting the Next Speaker

We will now turn to the card-sorting game and see to what extent Furhat is able to select the next speaker in a multiparty interaction using gaze (Skantze, Johansson, and Beskow 2015). Being able to shape

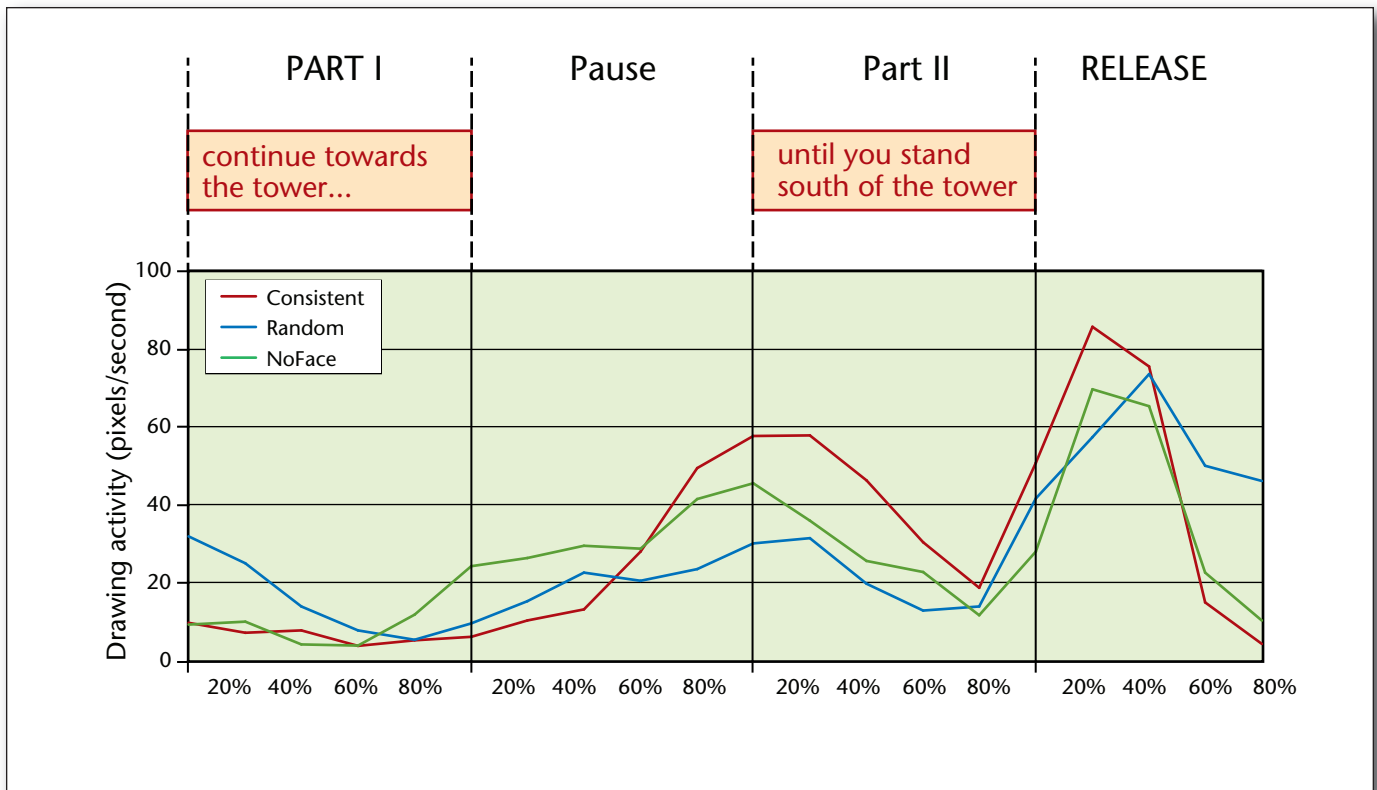


Figure 7. The Effect of Joint Attention on the Drawing Activity.

the interaction in this way could be important, for example, if it is desirable to involve both users in the interaction and balance their speaking time. To investigate this, we systematically varied the target of Furhat's gaze when asking questions during the museum exhibition, either toward both users (looking back and forth between them), toward the previous speaker (the one who spoke last), or toward the other speaker. An analysis of 2454 questions posed by Furhat is shown in figure 8. Overall, when Furhat targeted one user, that person was most likely to take the turn. If Furhat looked at both of them, the previous speaker was more likely to continue than the other speaker. If Furhat looked at the speaker who did not speak last (Other), the addressee was even more inclined to take the turn than if Furhat looked at the Previous speaker. Thus, Furhat can indeed help to distribute the floor to both speakers. If we split these distributions depending on whether the addressee is actually looking back at Furhat (mutual gaze), we can see that this makes the addressee even more likely to respond. This suggests that it is important for the robot to actually monitor the user's attention and seek mutual gaze, in order to effectively hand over the turn. To put it in other words, addressee selection is also a coordinated activity.

Claiming the Floor

Finally, we will look at how turn-holding cues can be used by the robot to claim the floor. Of course, if the robot is ready to speak immediately after the previous turn, there might not be any need for special cues to indicate the start of a turn. However, in the card-sorting game, we used cloud-based speech recognizers that give a relatively high accuracy, but the process takes about a second to complete. This could easily result in confusion if the system does not clearly signal that it has detected that it was being addressed and is about to respond. If the user doesn't get any response, there is a risk that she will continue speaking just when the robot starts to respond. A similar phenomenon occurs in human-human interaction, where speakers handle processing delays by starting to speak without having a complete plan of what to say (Levelt 1989). In such situations, it is common to start the utterance with a turn-holding cue (see table 1), for example a filled pause ("uhm..."), to signal that a response is about to come.

To investigate the effectiveness of such cues, we systematically experimented with different turn-holding cues for claiming the floor during the museum exhibition (Skantze, Johansson, and Beskow 2015). Figure 9 shows a schematic example where the user asks a question, and the system is not ready to respond until about 1300 ms later. Depending on

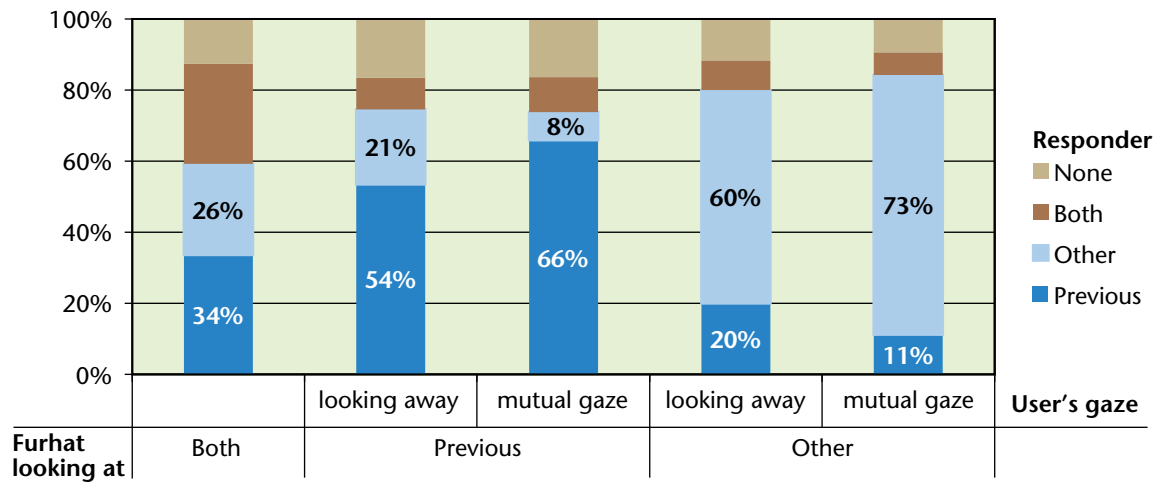


Figure 8. The Next Speaker in the Interaction, Depending on Furhat's and Users' Gaze.

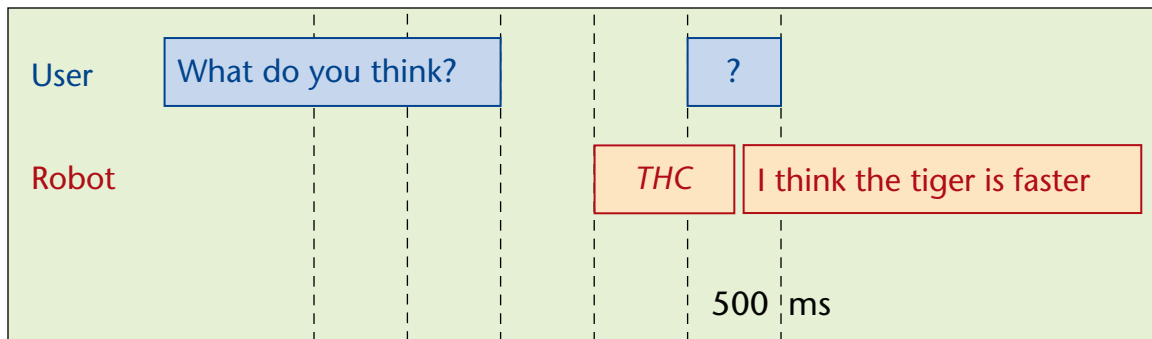


Figure 9. How the System Can Use Turn-Holding Cues to Claim the Floor when the Response Is Delayed.

We want to avoid that the user continue speaking (marked with "?").

the turn-holding cue (THC) used, we can expect different probabilities for the user to continue speaking in the window marked with "?" (which we want to minimize). This way, we can measure the effectiveness of different cues. As discussed above, humans often gaze away to hold the floor. This behavior was randomly used as a cue in 50 percent of the cases, and was contrasted with keeping the gaze toward the user in the other cases. In combination with this, we randomly selected between four different other cues:

(1) filled pause ("uhm..."), (2) a short breath, (3) smile, or (4) none of these. The breath was done by opening Furhat's mouth a bit and playing a recorded inhalation sound. Although smiling is not an obvious turn-holding cue, the purpose of the smile was to silently signal that the system somehow had reacted to the user's utterance. Thus, in total, we used 8 (2 x 4) different combinations of cues. In total, 991 such instances were analyzed, and the result is shown in figure 10. As can be seen, there is a main effect of gaz-

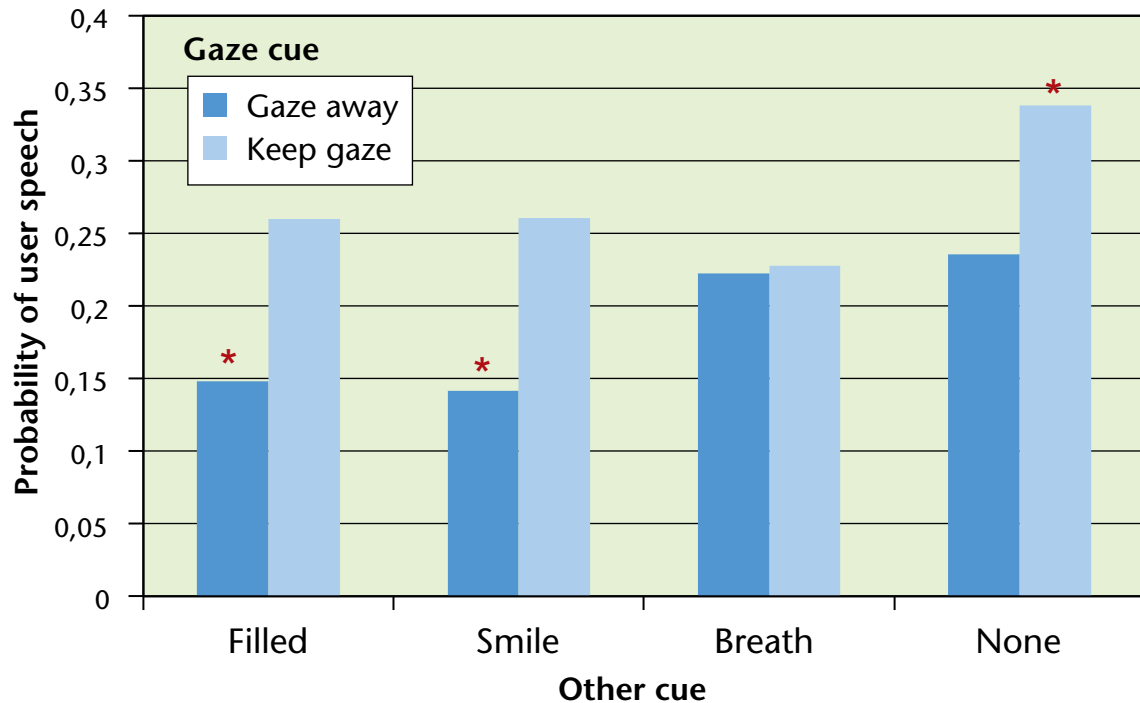


Figure 10. Probability That the User will Continue Speaking Depending on the Turn-Holding Cue(s) Used.

Significant deviations from overall distribution are marked with (*).

ing away, as expected. Looking at the other cues, they were all significantly more inhibiting than no cue. However, the strongest effect is achieved by combining cues, where a filled pause or a smile in combination with gazing away gives a significantly lower probability that the user will continue speaking (less than 15 percent), and no cues give a significantly higher probability (33.8 percent). This indicates that the cues humans use for coordinating turn-taking can be transferred to a humanlike robot and have similar effects. The fact that different combinations of cues can achieve the same effect is encouraging, since this makes it possible to use a more varied behavior in the robot.

Conclusions and Future Directions

Taken together, these results show that coordination is an important aspect of human-robot interaction, and that this coordination should be modeled on a much finer time scale than a simple turn-by-turn protocol. From studies of human-human interaction, we know that this coordination is achieved through subtle multimodal cues in the voice and face, including

words, prosody, gaze, gestures, and facial expressions. Thus, if we want robots to take part in real-time coordination, the underlying system must not only be able to pick up these cues and model these aspects, but the robot must also be able to express them. This has to be taken into account in the design of the robot. It could be argued that this coordination could be achieved through other signals than the ones humans make use of, for example with a lamp blinking when the robot is listening (Funakoshi et al. 2010). However, I would argue that, if possible, it makes more sense to use cues that we humans already know how to process and (unconsciously and automatically) pay attention to. It is also more likely that we will be able to emotionally relate to a robot that exhibits humanlike behaviors than one with more machinelike behavior. Of course, there is always a risk that the uncanny valley³ could give an opposite effect, but so far we have not seen many signs of that with Furhat, possibly because of its slightly cartoonish appearance.

Our results show that users in interaction with a humanlike robot make use of the same coordination signals typically found in studies on human-human

interaction. Thus, they do select the next speaker using gaze, and their prosody reflects whether they want to yield the turn or not. We have shown that the system can detect these cues by automatic means and combine them into turn-taking decisions with a fairly high accuracy. But we have also found interesting new correlations between how short feedback utterances reflect their temporal relationship with task progression (drawing the route on the map). Thus, the automatic extraction of features and fine-grained temporal resolution in our setups allow us to make new findings that we haven't seen in the literature on human-human interaction before.

We have also seen that humans react naturally to humanlike coordination signals when used by a robot, without being given any special instructions. They follow the gaze of the robot to disambiguate referring expressions, they conform when the robot selects the next speaker using gaze, and they naturally interpret subtle turn-holding cues, such as gaze aversion, breathing, facial gestures, and hesitation sounds in an expected way. These things are very important, if the robot should be able to shape the interaction, and avoid confusion.

A general finding that is consistent with the literature on human-human turn-taking is that face-to-face interaction gives a rich source of multimodal turn-taking cues, and that different combinations of turn-taking cues can achieve a similar effect. This is beneficial for human-robot interaction, since it allows for more robust interpretation of turn-taking cues (if there are uncertainties in some modalities), and allows the system to display a more varied behavior, while still achieving the same effect.

There are several ways in which we plan to further advance with this research program. When it comes to interpreting coordination signals, we have shown that this can be learned from data using an annotated corpus. However, we think that it is important that this could also be learned directly from the interaction, without the need for annotation, both because annotation is time consuming, but also because users might have very different behaviors that the robot should adapt to. By monitoring how the robot's turn-taking behavior results in either smooth turn-taking or in interaction problems (such as overlapping speech or long gaps), the robot can get automatic feedback on its behavior and thereby train the turn-taking model automatically in an unsupervised (or implicitly supervised) fashion, without the need for manual annotation. If several humans are interacting with the robot, it should also be possible to further improve the turn-taking model by observing where the humans take the turn when talking to each other.

Finally, we should add that the standard model of turn-taking by Sacks, Schegloff, and Jefferson (1974) has been challenged by other researchers, who argue that speakers do not always try to minimize gaps and

overlaps, but that the criteria for successful interaction is highly dependent on the kind of interaction taking place (O'Connell, Kowal, and Kaltenbacher 1990). In this view, overlaps do not always pose problems for humans; rather they could lead to a more efficient and engaging interaction. Thus, it is possible that robots should not necessarily always avoid overlaps. This view poses new challenges to our model, since it would require a more continuous decision of when to take the turn, rather than after each IPU. If we want such behavior to be learned online (as outlined above), we would also need to come up with new (measurable) criteria for successful interaction, rather than just minimizing gaps and overlaps.

Acknowledgements

This article is intended to give a summary and synthesis of some of the findings from several studies. The author would like to thank the other contributors to these experiments: Anna Hjalmarsson, Martin Johansson, and Catharine Oertel. This research was supported by the Swedish research council (VR) projects Incremental Processing in Multimodal Conversational Systems (#2011-6237) and Coordination of Attention and Turn-taking in Situated Interaction (#2013-1403), led by Gabriel Skantze.

Notes

1. A video of the interaction can be seen at www.youtube.com/watch?v=5fhjuGu3d0I.
2. www.irstk.net.
3. The phenomenon that nearly (but not perfectly) human-like faces might be perceived as creepy (Mori 1970).

References

- Al Moubayed, S.; Edlund, J.; and Beskow, J. 2012. Taming Mona Lisa: Communicating Gaze Faithfully in 2D and 3D Facial Projections. *ACM Transactions on Interactive Intelligent Systems* 1(2): 25. [dx.doi.org/10.1145/2070719.2070724](https://doi.org/10.1145/2070719.2070724)
- Al Moubayed, S.; Skantze, G.; and Beskow, J. 2013. The Furhat Back-Projected Humanoid Head — Lip Reading, Gaze and Multiparty Interaction. *International Journal of Humanoid Robotics* 10(1). [dx.doi.org/10.1142/S0219843613500059](https://doi.org/10.1142/S0219843613500059)
- Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E. F. 2000. *Embodied Conversational Agents*. Boston, MA: The MIT Press.
- Clark, H. H. 1996. *Using Language*. Cambridge, UK: Cambridge University Press. [dx.doi.org/10.1017/CBO9780511620539](https://doi.org/10.1017/CBO9780511620539)
- Clark, H. H., and Krych, M. A. 2004. Speaking While Monitoring Addressees for Understanding. *Journal of Memory and Language* 50(1): 62–81. [dx.doi.org/10.1016/j.jml.2003.08.004](https://doi.org/10.1016/j.jml.2003.08.004)
- Duncan, S. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* 23(2): 283–292. [dx.doi.org/10.1037/h0033031](https://doi.org/10.1037/h0033031)
- Funakoshi, K.; Nakano, M.; Kobayashi, K.; Komatsu, T.; and Yamada, S. 2010. Non-Humanlike Spoken Dialogue: A Design Perspective. In *Proceedings of the SIGDIAL 2010 Conference, The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 176–184. Stroudsburg, PA: Associ-

ation for Computational Linguistics.

Gravano, A., and Hirschberg, J. 2011. Turn-Taking Cues in Task-Oriented Dialogue. *Computer Speech and Language* 25(3): 601–634. dx.doi.org/10.1016/j.csl.2010.10.003

Heldner, M., and Edlund, J. 2010. Pauses, Gaps and Overlaps in Conversations. *Journal of Phonetics* 38(4): 555–568. dx.doi.org/10.1016/j.wocn.2010.08.002

Ishii, R., Otsuka, K., Kumano, S., and Yamato, J. 2014. Analysis of Respiration for Prediction of “Who Will Be Next Speaker and When?” in Multi-Party Meetings. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction* (ICMI 2014), 18–25. New York: Association for Computing Machinery. dx.doi.org/10.1145/2663204.2663271

Johansson, M., and Skantze, G. 2015. Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/w15-4642

Katzenmaier, M.; Stiefelhagen, R.; Schultz, T.; Rogina, I.; and Waibel, A. 2004. Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In *Proceedings of the 6th ACM International Conference on Multimodal Interaction* (ICMI 2004), 144–151. New York: Association for Computing Machinery. dx.doi.org/10.1145/1027933.1027959

Kendon, A. 1967. Some Functions of Gaze Direction in Social Interaction. *Acta Psychologica* 26: 22–63. dx.doi.org/10.1016/0001-6918(67)90005

Koiso, H.; Horiuchi, Y.; Tutiya, S.; Ichikawa, A.; and Den, Y. 1998. An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. *Language and Speech* 41(3–4): 295–321.

Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Mori, M. (1970). The Uncanny Valley. *Energy* 7(4): 33–35.

O’Connell, D. C.; Kowal, S.; and Kaltenbacher, E. 1990. Turn-Taking: A Critical Analysis of the Research Tradition. *Journal of Psycholinguistic Research* 19(6): 345–373. dx.doi.org/10.1007/BF01068884


Sacks, H.; Schegloff, E.; and Jefferson, G. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50(4): 696–735. dx.doi.org/10.1353/lan.1974.0010

Skantze, G., and Al Moubayed, S. 2012. IrisTK: A Statechart-Based Toolkit for Multi-Party Face-to-Face Interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (ICMI 2012). New York: Association for Computing Machinery. dx.doi.org/10.1145/2388676.2388698

Skantze, G.; Hjalmarsson, A.; and Oertel, C. 2014. Turn-Taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication* 65(Nov.–Dec.): 50–66. dx.doi.org/10.1016/j.specom.2014.05.005

Skantze, G.; Johansson, M.; and Beskow, J. 2015. Exploring Turn-taking Cues in Multi-Party Human-Robot Discussions About Objects. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI 2015). New York: Association for Computing Machinery. dx.doi.org/10.1145/2818346.2820749

Turing, A. M. 1950. *Computing Machinery and Intelligence*.




STITCH FIX

algorithms for true personalization

Combining art and science to transform the way people

find what they love



```

def style(ver): {...}
for i in j:
    curate(s_i, s_j)
    sel.append(s_j)
return sel

```

Mind 59(236): 433–460. dx.doi.org/10.1093/mind/LIX.236.433

Vinyals, O.; Bohus, D.; and Caruana, R. 2012. Learning Speaker, Addressee and Overlap Detection Models from Multimodal Streams. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (ICMI 2012), 417–424. New York: Association for Computing Machinery. dx.doi.org/10.1145/2388676.2388770

Weizenbaum, J. 1966. ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the Association for Computing Machinery* 9(1): 36–45. dx.doi.org/10.1145/365153.365168

Gabriel Skantze is an associate professor in speech technology at the Department of Speech Music and Hearing at KTH (Royal Institute of Technology), Stockholm, Sweden. He has a M.Sc. in cognitive science and a Ph.D. in speech technology. His primary research interests are in multimodal real-time dialogue processing, speech communication, and human-robot interaction, and he is currently leading several research projects in these areas. He is currently serving on the scientific advisory board for SIGdial, the ACL Special Interest Group on Discourse and Dialogue. He is the main architect and developer of IrisTK, an open-source software toolkit for human-robot interaction. He is also cofounder of the company Furhat Robotics.