Computer Speech & Language 67 (2021) 101178

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# Turn-taking in Conversational Systems and Human-Robot Interaction: A Review

Gabriel Skantze

*Department of Speech Music and Hearing, KTH, Sweden*

ARTICLE INFO

*Article History:*
Received 25 August 2020
Accepted 4 December 2020
Available online 16 December 2020

*Keywords:*
Turn-taking
Dialogue systems
Social robotics
Prosody
Gaze

ABSTRACT

The taking of turns is a fundamental aspect of dialogue. Since it is difficult to speak and listen at the same time, the participants need to coordinate who is currently speaking and when the next person can start to speak. Humans are very good at this coordination, and typically achieve fluent turn-taking with very small gaps and little overlap. Conversational systems (including voice assistants and social robots), on the other hand, typically have problems with frequent interruptions and long response delays, which has called for a substantial body of research on how to improve turn-taking in conversational systems. In this review article, we provide an overview of this research and give directions for future research. First, we provide a theoretical background of the linguistic research tradition on turn-taking and some of the fundamental concepts in theories of turn-taking. We also provide an extensive review of multi-modal cues (including verbal cues, prosody, breathing, gaze and gestures) that have been found to facilitate the coordination of turn-taking in human-human interaction, and which can be utilised for turn-taking in conversational systems. After this, we review work that has been done on modelling turn-taking, including end-of-turn detection, handling of user interruptions, generation of turn-taking cues, and multi-party human-robot interaction. Finally, we identify key areas where more research is needed to achieve fluent turn-taking in spoken interaction between man and machine.

## 1. Introduction

Many human social activities require some kind of turn-taking protocol, which determines the order in which different actions are supposed to take place, and by whom. This is obvious when, for example, playing a game of chess (where the protocol is very simple), 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. How this turn-taking is coordinated has been studied during the past decades in different scientific disciplines, including linguistics, phonetics, neuropsychology, and sociology.

Turn-taking is however not only a concern for those trying to understand human communication. As conversational systems (in various forms) are becoming ubiquitous, it is clear that turn-taking is still not handled very well in those systems. They often tend to interrupt the user or have very long response delays, there is little timely feedback, and the flow of the conversation feels stilted. Thus, modelling turn-taking in conversational systems is still very much an area of active research. In this review article, we will give an overview of findings from studies of turn-taking in human-human conversation, outline the state-of-the art in

*E-mail address:* gabriel@speech.kth.se

https://doi.org/10.1016/j.csl.2020.101178

modelling turn-taking in conversational systems and human-robot interaction, and draw conclusions about future directions for this research field.

The first dialogue systems, such as Eliza (Weizenbaum, 1966), were text-based, and in these turn shifts were clearly indicated by the user pushing the “send” button. The same goes for the chatbots still used today, on for example websites or in messaging apps. Similarly, some speech-based systems have relied on push-to-talk mechanisms for managing turn-taking (e.g., Hemphill et al. 1990; Traum et al. 2007; 2012). Push-to-talk is however not very convenient if the hands are busy or when talking over distance (for example with a robot or a smart speaker). There is also a risk that the user forgets to push the button, pushes the button after starting to speak, or releases it before the utterance is complete (Skantze and Gustafson, 2009). Such explicit signals are of course sometimes used in human-human interaction as well, especially in simplex channel settings, for example the “over” signal used in walkie-talkie interactions. Another example of explicit turn-taking signalling are the wake-words used in today’s smart speakers and voice assistants, such as “Hey Siri” or “Alexa” (Gao et al., 2020). The wake-word gives a clear cue that the user wants to initiate a turn (although the end of the turn has to be detected by other means). The wake-word also helps to identify the addressee of the utterance (as being the voice assistant and not some other person), as well as allowing the user to barge-in more easily. However, while they can be effective, the use of explicit cues is not very convenient and often leads to an interaction that is less “conversational” (Woodruff and Aoki, 2003). In most conversational settings, we manage turn-taking efficiently without thinking about how it is actually accomplished.

When not using explicit cues, spoken dialogue systems have traditionally detected the end of the user’s turn by a certain amount of silence. Silence, however, is not a very good indicator of turn-endings, since users might pause within a turn and thereby unintentionally trigger a response. This can be mitigated by increasing the silence threshold, but this will then lead to more sluggish responses. Studies of human-human conversation have found that pauses within turns are on average longer than gaps between turns (Brady, 1968; Ten Bosch et al., 2005; Edlund and Heldner, 2005), so silence is clearly not the main signal for humans to switch turns.

The signals (or cues) by which speakers coordinate their turn-taking have been studied extensively, and have been found across different modalities, including verbal cues, prosody, breathing, eye gaze and gestures, that we will explore in depth in Section 3 of this review. Thus, apart from the auditory channel, the visual channel (the face and body) are also important for turn-taking. Therefore, conversational systems that involve virtual or physical agents are also very interesting from a turn-taking perspective, as they provide a wider repertoire of turn-taking cues, and will also be covered in this review. Human-robot interaction (HRI) is especially interesting, as the situated nature of the interaction more easily allows for multi-party conversations, where turn-taking is an even more complex phenomenon.

This review article is organised as follows. We will start with a brief review of the linguistic research tradition on turn-taking and introduce some fundamental concepts (Section 2). Then we will review the cues that have been found to facilitate turn-taking across different modalities (Section 3). After this, we will review four main aspects of turn-taking in conversational systems that have so far attracted a considerable amount of research:

- How can the system identify appropriate places to take the turn or to produce a backchannel? (Section 4)
- How can the system handle interruptions, overlaps and backchannels from the user? (Section 5)
- How can the system generate turn-taking signals that help the user to understand whether the floor is open or not? (Section 6)
- How can the system handle multi-party and situated interaction (which is common for human-robot interaction), where there might be several potential addresses of an utterance, and which might involve the manipulation of physical objects? (Section 7)

Finally, in Section 8, we will identify a couple of directions where we think more research is needed.

As most research on turn-taking has been done on English (especialy when it comes to computational modelling), this will also be reflected in this review, with some exceptions. If the language of study is not explicitly stated, the reader should assume it is English.

## 2. Fundamental concepts

An example of how tightly coordinated turn-taking can be is shown in Fig. 1, where one person (speaker A) is describing a route on a map to another person (speaker B). As can be seen, there are very small (or no) gaps between the turns. Already before the first question from A is complete, B starts to answer. This is truly remarkable, given that B not only has to interpret what is being said, but also figure out a response, and then start to articulate that response (Levelt, 1989).

One of the most influential early accounts of the organisation of turn-taking is the one proposed by Sacks et al. (1974). Their model (which applies to dyadic as well as multi-party interaction) is based on a set of basic observations. To start with, the organisation is not planned in advance, but has to be coordinated in a flexible manner as the dialogue evolves. Also, they note that “overwhelmingly one party talks at a time. [... ] Occurrences of more than one speaker at a time are common but brief [... ] Transitions (from one turn to the next) with no gap and no overlap are common. Together with transitions characterised by slight gap or slight overlap, they make up the vast majority of transitions” (p. 700).

Based on these observations, they propose that turn-taking can be analysed using units of speech called **turn-constructional units (TCU)**, which are stretches of speech from one speaker during which other participants assume the role as listeners. After

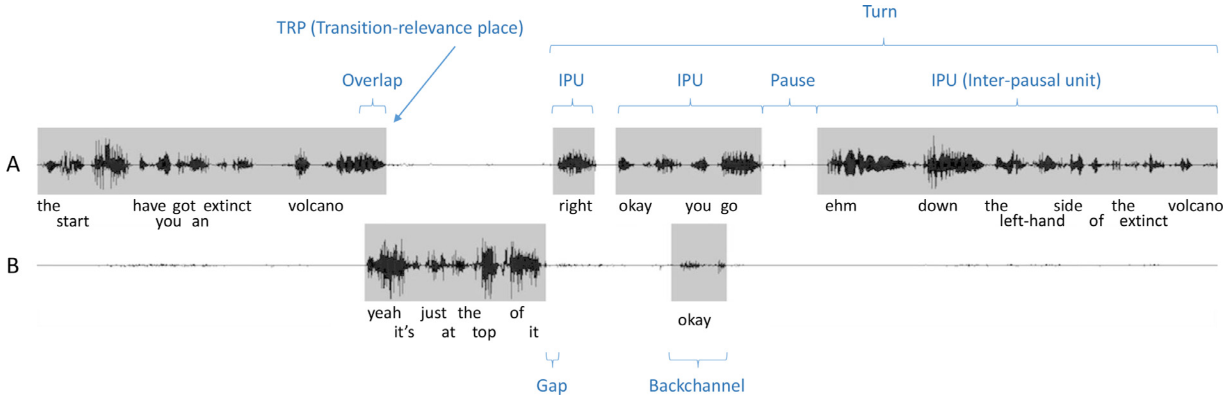

Turn
TRP (Transition-relevance place)
Overlap
IPU
IPU
Pause
IPU (Inter-pausal unit)
A
the start have got you an extinct volcano
right
okay you go
ehm down the left-hand side of the extinct volcano
B
yeah it's just at the top of it
okay
Gap
Backchannel

**Fig. 1.** Example of turn-taking from Map Task (Anderson et al., 1991).

each such unit, there is a **transition-relevant place (TRP)**, where a turn-shift can (but does not have to) occur according to the following rules:

1. The current speaker may select a next speaker (**other-select**), using for example gaze or an address term. In the case of dyadic conversation, this may default to the other speaker.
2. If the current speaker does not select a next speaker, then any participant can **self-select**. The first to start gains the turn.
3. If no other party self-selects, the current speaker may continue.

In order to identify TCUs and TRPs, researchers using speech technology have found it convenient to first segment the speech into **Inter-pausal units (IPUs)**, which are stretches of audio from one speaker without any silence exceeding a certain amount (such as 200ms). These can relatively easily be identified using voice activity detection (VAD). This was noted already by Brady (1965), who used this technique to analyse turn-taking patterns with automatic methods. A turn is then typically defined as a sequence of IPUs from a speaker, which are not interrupted by IPUs from another speaker. A possible exception are very short IPUs (like “mhm”) that might occur without the intention of “taking the turn”. We will discuss such **backchannels** further down. Silence between two IPUs within the same speaker are often referred to as **pauses**, whereas silence between IPUs from different speakers are called **gaps**. These concepts are illustrated in Fig. 1.

By operationalising turn-taking using IPUs, it possible to analyse turn-taking patterns and statistics in larger corpora with automatic methods (e.g. Brady 1968; Ten Bosch et al. 2005; Heldner and Edlund 2010; Levinson and Torreira 2015). One example of this is the histogram of turn-taking latency shown in Fig. 2. As can be seen, even if gaps and overlaps are common, humans are

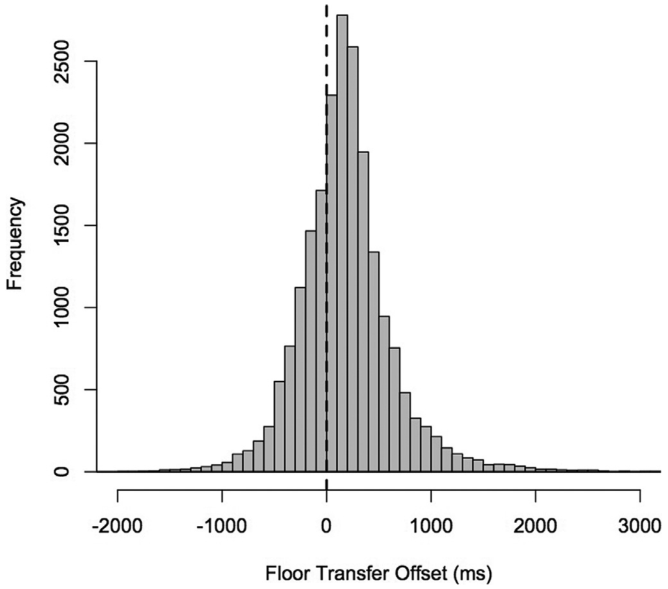


**Fig. 2.** Turn-taking latency in the Switchboard corpus (Godfrey et al., 1995), as calculated and visualised by Levinson and Torreira (2015). Negative latencies represent overlaps and positive latencies gaps.

**Table 1**
Typical turn-final cues found in studies of English conversation.

| | Turn-yielding cues | Turn-holding cues |
|---|---|---|
| Verbal | Syntactically complete | Syntactically incomplete, |
| Filled pause | | |
| Prosody | Rising or falling pitch | Flat pitch |
| Lower intensity | Higher intensity | |
| Breathing | Breathe out | Breathe in |
| Gaze | Looking at addressee | Looking away |
| Gesture | Terminated | Non-terminated |

typically very good at keeping them short, often with just a 200ms gap (although reported medians vary across interaction styles and cultures, as discussed in Section 2.5). It is clear that a response time of 200ms is much shorter than the silence threshold commonly used in spoken dialogue systems (often around 700–1000ms). Thus, there has to be some other mechanisms in place that help the listener to take turns with such small gaps, while minimising the amounts of overlaps.

### 2.1. Turn-taking cues

Several studies have investigated cues that can be found at the end of IPUs, which could be used by the listener to distinguish TRPs (**turn-yielding cues**) from non-TRPs (**turn-holding cues**). One of the first to systematically study these cues was Duncan and his associates (Duncan, 1972; 1974; Duncan and Fiske, 1977). By analysing face-to-face (American English, dyadic) conversations, they identified several multi-modal cues that signal turn-completion, including phrase-final intonation, termination of hand gesticulation, and completion of a grammatical clause. It is important to stress that these cues do not have a definite effect on the recipient, and that turn-taking is highly optional. However, what they found was that these cues had an additive effect: the listener was more likely to take the turn as the number of turn-yielding cues increased. These studies have later been followed by a large number of studies that have investigated the effects of individual cues, as well as the effects of combining them, using larger datasets, automatic methods, and more thorough statistical analyses (e.g. Koiso et al. 1998; Gravano and Hirschberg 2011; Hjalmarsson 2011). In general, these studies tend to confirm the finding that turn-taking cues are additive, even if there is also a considerable amount of redundancy. A summary of typical cues found in studies of turn-taking is listed in Table 1. We will provide an in-depth review of these cues in Section 3.

A complicating factor in these kind of analyses is that TRPs (in the sense proposed by Sacks et al.) cannot be directly observed in data. We can only observe actual turn-shifts, which might thus be considered to be a subset of TRPs. Furthermore, turn-shifts might also occur where there is no TRP (for example if the listener has something very important to say), as speakers are of course free to “break the rules”, just like they are free to produce “non-grammatical” utterances. Given this, some researchers have questioned whether it is meaningful to talk about turn-taking “rules” at all (O’Connell et al., 1990). In any case, it seems clear that turn-shifts are more likely to occur at certain places than others. Thus, we would like to propose that TRP should not be regarded as a binary notion, but rather as the probability of a speaker shift in a certain context (for example given a certain set of cues).

In addition to the signals produced by the current speaker, it is also relevant to consider the signals produced by the next speaker to indicate her willingness to take the turn. Duncan (1974) argues that there are certain signals which show that the next speaker wants to take the turn (when self-selecting or when accepting an offered turn), which we will refer to as **turn-initial cues**. These cues include things like gazing away, inhaling or initiating a gesture (ibid.). Thus, they are similar to the turn-holding cues discussed above, although they are produced before the onset, or just at the start, of a turn. The turn-initial cue can also help to differentiate attempts to take the turn from backchannels, which will be discussed in Section 2.3 below.

### 2.2. Reaction vs. prediction

While the turn-taking cues outlined above provide some explanation as to how the listener can differentiate pauses from gaps, they cannot provide a full account for how turn-taking is coordinated. If the listener only focused on cues at the end of the turn, it would not really be plausible to find a typical response time of 200ms (as seen Fig. 2). This would not give the listener enough time to react to the cue, prepare a response and start speaking. According to psycholinguistic estimates, the response time would rather be around 600-1500ms (Levinson and Torreira, 2015). This has led many researchers to conclude that there must be some sort of prediction mechanism involved (Sacks et al., 1974; Levinson and Torreira, 2015; Garrod and Pickering, 2015; Ward, 2019). This is even more evident when considering the fairly large proportion of turn-shifts that occur without any gap at all, i.e., before the IPU has even been completed (as seen in Fig. 2). An example of such a turn-shift was shown in Fig. 1. Well before the first question is complete (before the final word “volcano” is spoken), speaker B must predict that the turn is about to end, what kind of dialogue act is being produced (a question), as well as the final word in the question, in order to prepare a meaningful answer.

This predictive view is sometimes contrasted with the “Signalling approach” (or “reactive” view) to turn-taking which focuses on the cues found at the end of the turn. However, even though the predictive and reactive accounts of turn-taking are sometimes portrayed as two opposing views, it seems like most researchers today acknowledge the need for both prediction mechanisms (which allows the other speaker to prepare a response), as well as some turn-final cues that can confirm that the turn is actually complete and yielded (Heldner and Edlund, 2010; Levinson and Torreira, 2015). Ward (2019) makes an analogy with the "Ready-Set-Go!" signal used when starting a sprint.

The mechanisms used for prediction are more complex to study and harder to identify than the signals found at the end of the turn. Sacks et al. (1974) argued that syntax and semantics were more important for prediction than for example intonation, as the completion of a syntactic unit is likely to be more predictable than a prosodic unit (more on that in Section 3). Garrod and Pickering (2015) argue that dialogue act prediction is an integral part of end-of-turn prediction. In their account, the addressee covertly imitates the speaker's utterance in order to determine the underlying intention of the upcoming utterance (in terms of content and timing). In addition, they argue that the speakers rely on "entrainment of low-frequency oscillations between speech envelope and brain", as originally proposed by Wilson and Wilson (2005).

### *2.3. Overlaps, backchannels and interruptions*

As we have seen, even though dialogue proceeds predominately with one speaker at a time, there is typically a considerable amount of overlap. It is important to note that overlap should not just be considered as "failed" turn-taking, as it often serves many important functions and contributes to a fluent interaction (Coates, 1994). A distinction can be made between **competitive** and **cooperative overlap**. In competitive overlaps, the two speakers are seen as competing for the turn, where one of them will have to "give up" the turn. In cooperative overlaps, they are seen as producing speech in a collaborative manner, and are not competing for the turn.

Schegloff (2000) makes a further distinction between four types of cooperative overlaps:

- **Terminal overlaps**: The listener predicts the end of the turn and starts speaking before it is completed, as illustrated in Fig. 1
- **Continuers** (or **Backchannels**): Brief, relatively soft vocalisations, such as "mm hm", "uh huh", or "yeah", produced by the listener to show continued attention and possibly aspects such as attitude and uncertainty (Ward, 2004). This phenomenon has been referred to as "backchannels" (Yngve, 1970), "listener responses" (Dittman and Llewellyn, 1967) and "accompaniment signals" (Kendon, 1967). In a face-to-face setting, backchannels can also be produced in the visual channel, for example by nodding or making facial expressions.
- **Conditional access to turn**: Cases where the listener helps the speaker to construct the turn and possibly bring it to completion. This might be to just fill in a name that the speaker has forgotten, or a longer sequence that is produced jointly. This is often referred to as **sentence completion** (Poesio and Rieser, 2010).
- **Choral talk**: Simultaneous production of speech. This includes laughter or greetings done in concert.

Backchannels hold a special status when it comes to turn-taking, since they are fairly frequent but still not typically considered to constitute a "turn". Thus, in automatic analyses of turn-taking based purely on VAD, they have to be accounted for somehow. Similar to how turn shifts are found more often after certain cues, the timing of backchannels is also thought to be associated with certain **backchannel-inviting cues**, where the speaker is looking for evidence of understanding from the listener (Clark, 1996). However, since the speaker typically intends to continue after the backchannel, these cues should look a bit different from turn-yielding cues. Thus, analogous to TRPs, we will use the term **backchannel-relevant places** (BRPs), which we will come back to in Section 3.

Since the listener does not intend to take the turn when producing a backchannel (or other forms of cooperative overlaps), it is important that the current speaker can differentiate these from attempts to take the turn. As discussed in Section 2.1 above (and in depth by Duncan 1974), the production of turn-initial cues can help in this regard. When it comes to the vocalisations of competitive or cooperative overlap, several studies have found that competitive overlaps have higher pitch and intensity (French and Local, 1983; Yang, 2001), as well as other acoustic properties (Oertel et al., 2012b; Truong, 2013). In a face-to-face setting, eyebrow movement and mouth opening can also be good predictors (Lee and Narayanan, 2010).

An interesting question is whether overlaps tend to occur at certain points rather than others. Dethlefs et al. (2016) investigated information density as a predictor for overlaps (a word with high information density is a word that is not very probable, given its context), as it might be easier to perceive overlapping speech if it is more predictable. They conducted a perception experiment, where overlapping speech from a dialogue system, in the form of synthesized interruptions and backchannels, were added to the speech from a user. When overlapping speech segments were inserted in regions with low information density, subjects rated them as more natural.

Unlike cooperative overlaps, competitive overlaps need some kind of resolution mechanism (to determine who should get the floor). According to Schegloff (2000), competitive overlapping talk is characterised by "hitches and perturbations" in the speech, which involves increase in intensity, higher pitch, change in pace, glottal stops, or repetitions. In his corpus analysis, most competitive overlaps were resolved (meaning that one of the participants gives up the turn) after one or two syllables.

It is important not to confuse overlap with **interruptions**. As pointed out by Bennett (1981), overlaps can be objectively identified in a corpus, whereas the notion of interruptions requires some form of interpretation, i.e., that some participant is violating the other participant's right to speak. It should also be stressed that interruptions is not the same thing as competitive overlap, as defined above, since interruptions can also occur without overlap, in the case a speaker makes a pause (completes an IPU without yielding the turn), and the other participant starts to speak (Gravano and Hirschberg, 2012). This makes it hard to automatically identify interruptions in a corpus purely based on VAD patterns, and a dialogue system might interrupt the user without any overlap involved (i.e., taking the turn after an IPU that is not a TRP). Based on manual annotation of interruptions in a task-oriented dialogue corpus, Gravano and Hirschberg (2012) found that the onset of non-overlapping IPUs labeled as interruptions had higher intensity, pitch level and speech rate.

### 2.4. Situated, multi-party interaction

When studying and modelling turn-taking, it is important to take the setting of the conversation into account. Intuitively, a face-to-face setting provides a richer repertoire of cues for coordinating turn-taking than a conversation over the phone and could therefore be expected to be more fluent. For example, seeing each others' faces allows us to perceive gaze direction and facial expressions. However, studies that compare spoken interaction in video meetings with voice-only interactions have not found any substantial differences when it comes to the coordination of turn-taking (O'Conaill et al., 1993; Sellen, 1995). But when comparing video conferences to physical face-to-face meetings, O'Conaill et al. (1993) found that the former had longer conversational turns, fewer overlaps and backchannels, as well as more formal mechanisms for shifting turns. Thus, it seems like the physical co-presence allows us to more easily pick up these visual cues and coordinate turn-taking. Especially in multi-party interaction, video conferences and animated agents on 2D displays are not very well suited for efficient turn-taking (Al Moubayed and Skantze, 2011). This is an important argument for why physical robots might provide better opportunities for social interaction compared to virtual agents or voice assistants (Skantze, 2016). However, even in dyadic interactions, the physical presence of a robot has been shown to be beneficial in for example language learning, compared to a virtual character on a screen (Leyzberg et al., 2012).

In dyadic interaction, it is always clear who is supposed to speak next when the turn is yielded. In multi-party interaction, on the other hand, this has to be coordinated somehow. Most people have experienced how problematic this can be when we lack clear signals for this coordination, such as in online meetings. As discussed in the beginning of Section 2, the basic model of turn-taking proposed by Sacks et al. (1974) also accounts for multi-party settings. In this model, the current speaker may select the next speaker, who then "has the right and is obliged" to take the next turn to speak, whereas no other participant is supposed to do so. If the current speaker does not select a next speaker, any participant has the opportunity to "self-select", or the current speaker may continue. This means that turn-taking in multi-party interaction should be even harder to predict, as the transitions are more optional.

If the current speaker selects the next speaker, a common signal is to gaze at the next speaker, which indicates the attention of the current speaker, but it is also possible to use other means, such as the addressee's name or a pointing gesture (Auer, 2018). However, as argued by Auer (2018), addressee selection and next-speaker selection is not always the same thing. It is for example possible to address a statement to several people (by alternatingly looking at them), while selecting a specific person as the next speaker (by finally looking at that person). Nevertheless, from a computational perspective, this distinction is often not made, and the problem of identifying the target of an utterance is often referred to as **addressee detection** (cf. Jovanovic et al. 2006; Katzenmaier et al. 2004; Vinyals et al. 2012), which we will come back to in Section 7.1.

In multi-party interaction, we can also distinguish several different roles of the (potential) participants. In dyadic interaction, each participant is currently either a *speaker* or *addressee* (the listener). However, multi-party interaction involves other types of listeners. Among those considered to be participants in the interaction, there might also be **side participants**, who are neither currently speaking or being addressed, but who might still take the turn at any point. But there might also be **overhearers** in the vicinity, who are not considered to be participants, such as **bystanders** (still openly present to the participants) and **eavesdroppers** (Goffman, 1979; Clark, 1996).

### 2.5. Cultural and developmental aspects

Even though the general patterns of turn-taking ("one speaker at a time") seem to be fairly generic across languages (Stivers et al., 2009), there are also notable differences when it comes to which specific cues are being used, and overall distributions. For example, Stivers et al. (2009) found mean gap length to vary substantially between 10 different languages, from 7ms in Japanese to 489ms in Danish. The frequency and placement of backchannels also seem to be different across languages. In a study of English, Mandarin and Japanese, Clancy et al. (1996) found that backchannels are most frequent in Japanese, fairly frequent in English, and least frequent in Mandarin. The placement of backchannels also seemed to be more aligned with grammatical completion points in English and Mandarin than in Japanese. When it comes to prosodic turn-taking cues, there are also differences, which will be discussed in Section 3.2.

The developmental aspect of turn-taking is important, since turn-taking is a skill that gets acquired and perfected relatively late in child language development. This learning begins by dyadic interaction with the caregiver, who is responsible for regulating most of the turn-taking (Ervin-Tripp, 1979). Later on, children also learn how to claim the floor in multi-party interaction, a skill that is mastered around the age of six. Even after that, children continue to learn how to take turns, and reduce gaps and overlaps. Whereas adults often take turns with very short gaps, children's gaps are typically longer – in some studies average gap length has been measured at 1.5–2 s (ibid.). This gap length shortens with age, as the child learns to better pick up turn-yielding cues, project the interlocutor's turn ending, and plan their own response (ibid.). Researchers have also compared child-child interaction with child-adult interaction (Martinez, 1987). Typically, children conversing with adults allow the adult to regulate the interaction (they are less inclined to self-select), and then pick up these regulatory behaviours and employ them when talking to other children.

## 3. Turn-taking cues

In this section, we will provide a more thorough review of the literature on turn-taking cues. As we will discuss in more depth in the next sections, to be able to engage in a more fluent interaction, a conversational system needs to be able to both understand and generate such cues.

Most of the more systematic studies on turn-taking cues have focused on cues found at the end of the turn. As discussed in Section 2.2, it is not plausible that turn-final cues will suffice to give a full account of how turn-taking is coordinated. However, turn-final cues are of course much easier to identify in a systematic manner than cues that appear earlier and which are used for projection of turn completion. Another problem with any corpus study is of course that it is hard to know whether cues correlated with certain behaviours are in fact used as signals by the speakers (correlation does not imply causation). To sort this out, corpus studies need to be complemented with controlled experiments. However, from a practical perspective, even if it turns out that a certain cue is not actually used by the human listener, it is still possible that it could be utilised by a conversational system, which has different computational constraints than the human brain.

### *3.1. Verbal cues: syntax, semantics and pragmatics*

As conversation ultimately progresses through the exchange of meaningful contributions (dialogue acts), the verbal aspect of spoken language (also referred to as “linguistic features”, i.e., the words spoken and the semantic and pragmatic information that can be derived from those) is likely very important for regulating turn shifts. The completion of a syntactic unit is intuitively a requirement for considering the turn as “done” (a TRP). For example, the phrase “I would like to order a...” is not syntactically complete, and the listener is likely to wait for the speaker to finish the sentence. Another argument for the prominence of verbal cues for turn-taking is the projectability of syntactic units, which might help to explain the precise turn-taking coordination with very small gaps discussed in Section 2.2 (Sacks et al., 1974; Ford and Thompson, 1996). Given the context in which the utterance is spoken, it might be possible to project the completion of the sentence and thereby predict roughly how soon it will come. Such predictions are also necessary for certain types of collaborative overlaps, such as the “choral” speech or sentence completions discussed in Section 2.3 above. Yet another argument for the importance of verbal cues is the fact that turn-taking is a skill that is perfected relatively late in child language development (as discussed in Section 2.5).

Ford and Thompson (1996) define an utterance to be **syntactically complete** if “in its discourse context, it could be interpreted as a complete clause, that is, with an overt or directly recoverable predicate” (p. 143). This includes “elliptical clauses, answers to questions, and backchannel responses”. Thus, syntactic completion (by this definition) does not have to be a complete sentence. Neither is a syntactic phrase (like a nominal phrase) necessarily syntactically complete. Syntactic completion is judged incrementally as the utterance unfolds. The following (made-up) example (from Ekstedt and Skantze, 2020) illustrates this notion, where / marks syntactic completion points:

(1) A: yesterday we met / in the park /
B: okay / when / will you meet / again /
A: tomorrow /

As can be seen, in this account, the turn-initial adverb of time “yesterday” is not syntactically complete (as there is no “overt or directly recoverable predicate”), whereas “tomorrow” is, which illustrates the dependence on the dialogue context.

As pointed out by Ford and Thompson (1996), while syntactic completion might be necessary for a TRP, it is not sufficient. They make a distinction between syntactic completion and **pragmatic completion**. The latter is defined as having “a final intonation contour and has to be interpretable as a complete conversational action within its specific sequential context” (p. 150). In their analysis, these are the points that constitute actual TRPs. There is no precise definition of what is considered to be a “complete conversational action”, and the annotator is likely to depend to a fair amount of common sense. In the example above, “okay when will you meet” could constitute a valid question in itself, but is unlikely, given the preceding context. Thus, it would be syntactically, but not pragmatically, complete. In the corpus analysis of Ford and Thompson (1996), about half of all syntactic completions were also pragmatic completions.

We think the notion of pragmatic completion is useful, but it is not entirely clear why Ford and Thompson (1996) include prosodic aspects in their definition of pragmatic completion. We argue it would be better to reserve the notion of pragmatic completion to the verbal aspect of speech. In this view, pragmatic completion points form a set of potential TRPs, to which other modalities (prosody, gaze, etc) add further constraints to derive actual TRPs, as will be discussed in the next sections. In many cases, pragmatic completion is not in itself sufficient for the listener to know when the speaker intends to yield the turn, as the following example shows:

(2) I would like a hamburger / with fries / and a milkshake /

Again, it is important to stress that TRPs do not automatically imply turn shifts. In the analysis of Ford and Thompson (1996), of all the TRPs identified through pragmatic completion (which in their account also included intonation), about half of them involved actual turn shifts. In many cases, it also seemed like the next speaker started to speak at an earlier TRP, while the first speaker further amended her utterance, resulting in a terminal overlap.

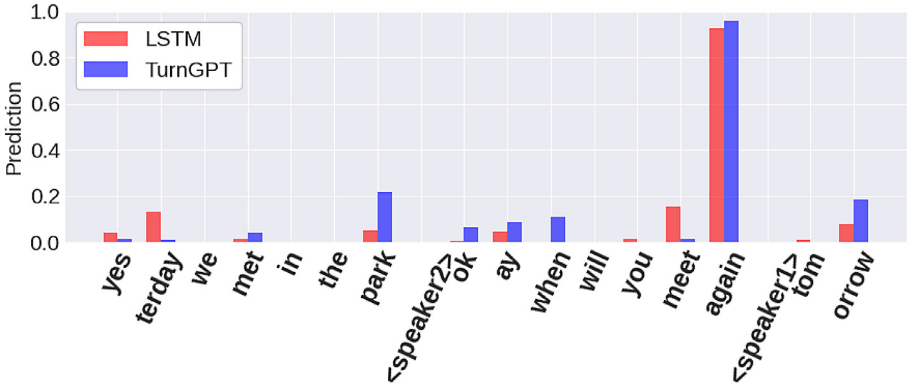


**Fig. 3.** Turn-shift probability, as predicted with the transformer-based model TurnGPT (vs. an LSTM model). Some words are split into sub-words. Figure from Ekstedt and Skantze (2020).

As we have seen, syntactic and pragmatic completion can be very hard to determine, especially since they often require the consideration of the preceding dialogue context. Most of the more sophisticated analyses of turn-taking in the Conversation Analysis (CA) tradition have therefore relied on manual annotation. This can be problematic, as it is hard to annotate pragmatic and syntactic completion without being influenced by the actual turn shifts in the corpus. This makes it challenging to study the role of verbal cues in turn-taking.

When incorporating syntax into computational models of turn-taking, much simpler operationalisations have typically been used, where the dialogue context is not considered at all. Several models have for example used the two final part-of-speech tags (Gravano and Hirschberg, 2011; Meena et al., 2014; Koiso et al., 1998). For instance, a syntactically complete phrase is more likely to end with a noun, but less likely to end with a conjunction or determiner. However, this can of course not account for the much more sophisticated notion of syntactic or pragmatic completion discussed above. More recent turn-taking models have used LSTMs to encode linguistic information, such as part-of-speech (Skantze, 2017b), words (Roddy et al., 2018a) or senones (Masumura et al., 2018).

Although several of these studies have found that linguistic information contribute to the performance (compared to for example only using prosody), the performance gain is perhaps not as big as what could be expected. One explanation for this could be the lack of proper modelling of the dialogue context. Recently, the use of stronger transformer-based language models for identifying TRPs has been proposed by Ekstedt and Skantze (2020), showing a substantially better performance than the use pf turn-final words or LSTMs. Further analyses of the models also showed that they indeed do utilise the context of the preceding turns. The model's predictions on example (1) above is shown in Fig. 3. As can be seen, unlike the LSTM, the model accurately predicts that a turn-shift is more likely after "tomorrow" than after "yesterday", given the context. This also illustrates how TRPs can be modelled using a more probabilistic (rather than binary) notion.

A common verbal turn-holding or turn-initial cue, are so-called **fillers** or **filled pauses**, such as "uh" or "um" (Ball, 1975). These often indicate that the speaker is thinking about what to say next. An interesting question is whether these vocalisations are simply a symptom of a hesitation in the speech production process, or whether they are used more deliberately, and whether different lexical and prosodic realisations of these fillers have different meanings. Clark and Fox Tree (2002) argue for the latter interpretation. In their view, speakers monitor their own speech plan for upcoming delays, and use fillers (as if they were words) to comment on these delays. Based on manual annotation of a speech corpus, they conclude that "uh" is used to signal a shorter delay, whereas "um" signals a longer delay. However, O'Connell and Kowal (2005) argue against such an interpretation. Using acoustic measurements to analyse a dialogue corpus, they did not find any relationship between the choice of "uh" or "um" and subsequent delays, suggesting that they do not have different meanings. Nevertheless, regardless of their intentional status, fillers still likely serve as important turn-holding and turn-initial cues for the listener.

### 3.2. Prosody

The role of prosody in turn-taking has the been subject of much interest and dispute. Prosody refers to the non-verbal aspects of speech, including intonation, loudness, speaking rate and timber. It has been found to serve many important functions in conversation, including the marking of prominence, syntactic disambiguation, attitudinal reactions, uncertainty, and topic shifts (Ward, 2019). As we saw in the discussion on pragmatic completion, Ford and Thompson (1996) included **intonation** in their definition of the term. When it comes to intonation, studies across various languages have found that level intonation (in the middle of the speaker's fundamental frequency range) near the end of an IPU tend to serve as a turn-holding cue, including English (Duncan, 1972; Local et al., 1986; Gravano and Hirschberg, 2011), German (Selting, 1996), Japanese (Koiso et al., 1998) and Swedish (Edlund and Heldner, 2005). Complementary to this, studies of English and Japanese have found that either rising or falling pitch can be found in turn-yielding contexts (Gravano and Hirschberg, 2011; Local et al., 1986; Koiso et al., 1998). However, studies of

Swedish have found that while falling pitch is a turn-yielding cue, rising pitch is not clearly associated with either turn-holds or turn-shifts (Edlund and Heldner, 2005; Hjalmarsson, 2011).

Gravano and Hirschberg (2011) also looked at **intensity** in English dialogue, and found that speakers tend to lower their voices when approaching potential turn boundaries, whereas turn-internal pauses had a higher intensity. Similar patterns were found by Koiso et al. (1998) in Japanese dialogue, where low or decreasing energy was associated with turn change, and non-decreasing energy was associated with turn hold.

Regarding **duration** and speaking rate, findings seem to be mixed. Duncan (1972) found a "drawl on the final syllable or on the stressed syllable of a terminal clause" to be a turn-yielding cue (in English). This is also in line with the findings of Local et al. (1986). However, Gravano and Hirschberg (2011) found that final lengthening tends to occur at all phrase final positions, not just at turn endings. If anything, final lengthening seemed to be more prominent in turn-medial IPUs than in turn-final ones. In an analysis of Japanese task-oriented dialogue, Koiso et al. (1998) also found longer duration to be associated with turn hold. In a listening task where participants were asked to predict turn-shifts in Swedish, Hjalmarsson (2011) did not find final lengthening to result in any consistent predictions.

Gravano and Hirschberg (2011) also examined **voice quality**, as measured through jitter, shimmer and noise-to-harmonics ratio (NHR), and found these acoustic features to potentially serve as turn-taking cues. According to Ward (2019), **creaky voice** is also commonly found before turn-shifts.

Several studies have also examined how prosody can serve as a backchannel-inviting cue, and how this differs from turn-yielding cues. Ward (1996) investigated a corpus of Japanese conversations to find predictive cues for backchannels. He found that backchannels tended to come about 200ms after a region of low pitch. Gravano and Hirschberg (2011), on the other hand, found that IPUs immediately preceding backchannels showed a clear tendency towards a final rising intonation, as well as higher intensity (i.e., opposite to what they found to be a turn-yielding cue). These somewhat contradictory findings, could perhaps be explained by language differences (Japanese vs. American English), or the fact that Ward (1996) looked at overlapping backchannels in unconstrained conversations, whereas Gravano and Hirschberg (2011) looked at backchannels coming after an IPU in task-oriented dialogue. In an experiment where participants were given the task of dictating number sequences to each other, Skantze and Schlangen (2009) found that they segmented the longer number sequence into "installments" (Clark, 1996), where each installment ended with a rising pitch, which seemed to invite a brief feedback from the listener (usually something similar to a backchannel, but potentially also clarification requests), and then ended the full sequence with falling pitch.

As discussed earlier, since turn-final cues (by themselves) cannot explain rapid turn-shifts, it is not clear what role these prosodic features have for turn-taking. Ward (2019) goes so far as to call the importance of turn-final prosody for turn-taking a "popular myth", and argues that "for most turn exchanges, final prosody can play no role in the turn-start decision" (p. 145). However, he also notes that "turn-final prosody might still be decisive if the speaker has a response pre-prepared and only needs to decide whether to deploy it or not" (p. 210). Heldner and Edlund (2010) argue that even though many turn-shifts occur with very little gap (or even overlap), which might require prediction, there are still a considerable amount of turn-shifts that have a longer gap, where prosody could still play an important role.

It is also not clear to what extent prosody provides additional information compared to verbal cues, or if they are redundant. In an experiment by de Ruiter et al. (2006), subjects were asked to listen to a conversation and press a button when they anticipated a turn ending. The speech signal was manipulated to either flatten the intonational contour, or to remove verbal information by low-pass filtering. The results showed that the absence of intonational information did not reduce the subjects' prediction performance significantly, but that the subjects' performance deteriorated significantly in the absence of verbal information. From this, they concluded that verbal information is crucial for end-of-turn prediction, but that intonational information is neither necessary nor sufficient.

As we saw in the discussion on verbal cues in Section 3.1 above, Ford and Thompson (1996) included the requirement of a "final intonation contour" in their definition of pragmatic completion. This is a more complex (and subjective) notion than the prosodic features discussed above, and involves a longer prosodic gesture over an intonational unit, but it can to some extent be related to the falling (or rising) final pitch others have identified as turn-yielding. Thus, in their account, prosody helps to filter our syntactically complete units which are also pragmatically complete (and thereby TRPs).

If prosody is specifically useful in syntactically ambiguous places, this can help to explain the findings of de Ruiter et al. (2006) mentioned above, as the stimuli used might not have involved such ambiguities. Indeed, Bögels and Torreira (2015) performed a similar experiment, but selected the stimuli so that they contained several pragmatic completion points, and where the intonation phrase boundary provided additional cues to whether they were actual TRPs. They found that subjects made better predictions when the intonation was intact.

Taken together, these studies indicate that prosody can play an important role in cases where pragmatic completion in itself is not sufficient. They do not say, however, how common these type of ambiguities are, and how important prosody is overall for turn-taking. One way of addressing this question is to look at turn-taking prediction models and see how much prosodic features contribute to predictive performance, compared to verbal features. Koiso et al. (1998) investigated the relationship between syntax and prosody for predicting turn shifts in task-oriented Japanese dialogue. Using a C4.5 decision tree, they explored how the predictive power of the model changed when various syntactic and prosodic features were added or removed. The results showed that some of the individual syntactic features had very strong contributions, which was not the case of individual prosodic features. However, if all prosodic features were removed from the model, the performance dropped as much as when removing all syntactic features. In a similar way, Gravano and Hirschberg (2011) investigated American English task-oriented dialogue and trained a multiple logistic regression model to predict turn-shifts. The most important cue in their model was textual completion, followed by voice quality, speaking rate, intensity level, pitch level and IPU duration. However, when using all features, they did not find any significant contribution of intonation to the general predictive power of the model. One caveat in these type of studies is of course that these models do not encode and use prosody and verbal aspects in the same way as humans; especially pragmatic aspects are virtually non-existing, as we saw in Section 3.1. Therefore, the general contribution of verbal and prosodic cues to turn-taking is still very much an open question.

Regardless of the role of prosody in turn-taking between humans, prosody might provide important cues from the perspective of a conversational system. Since conversational systems do not have the same computational/cognitive constraints and might not have to prepare the response in advance to the same extent as humans, they could make use of turn-final cues to a larger extent. Also, whereas prosodic features can be extracted fairly reliably in a continuous fashion (Eyben et al., 2010), verbal features rely on an Automatic Speech Recogniser (ASR), which introduces a certain delay and ASR errors, as will be further discussed in Section 4.2 below. For these reasons, prosody might be more important for conversational systems than for humans. However, as we have seen, the coordinative functions of prosody seem to vary somewhat between languages and settings.

### 3.3. *Breathing*

Breathing is intuitively linked to turn-taking, as we typically breathe in before starting to speak, which means that an (audible and/or visible) in-breath might serve as a cue that the speaker intends to speak in the near future. In a study on breathing in conversation, McFarland (2001) found increased expiratory duration before speech onset at turn-shifts, which may reflect the preparation of the respiratory system for speech production.

Rochet-Capellan and Fuchs (2014) also investigated the role of breathing as a coordination cue. They found no global relationship between breathing and turn-taking rates and no signs of general breathing synchronisation (temporal alignment of breathing) between the participants. However, they did find a local relationship between turn-taking and breathing. When turn-taking was successful, speech onset was generally well timed with inhalation events. When a participant took a breath and tried to take the turn but failed, they shortened their breathing cycle. At turn-holds, the speaker also inhaled (although less so than at the beginning of the turn), indicating that they would like to continue speaking.

Torreira et al. (2015) examined inbreaths occurring right before answering a question, and found that they typically begin briefly after the end of questions. These inbreaths are also associated with substantially delayed answers (576ms vs. 100ms for answers not preceded by an inbreath), as well as longer answers. This indicates that breathing is linked to the planning of the response, and also means that breathing could be regarded as a turn-initial cue, showing that the next speaker has detected a turn-end and is preparing a response.

Ishii et al. (2014) examined breathing in multi-party interactions and found that when a speaker is holding the turn, she inhales more rapidly and quickly than when yielding the turn. The speaker who is about to take the turn tends to take a deeper breath compared to listeners who are not about to speak. Thus, it is possible that breathing helps to coordinate self-selection.

Most studies on breathing have used various forms of invasive measuring equipment, such as elastic strain gauges that measure movements of the rib cage (McFarland, 2001). However, for breathing to play a role as a cue for coordination in conversation, it needs to be audible and/or visible to the other participants. Włodarczak and Heldner (2016) examined acoustic inhalation intensity (as picked by a microphone) as a cue to speech initiation and found inhalations preceding speech to be louder than those in tidal breathing and before backchannels. However, while judges have been shown to be able perceive respiratory pauses in speech (Wang et al., 2012), it is not known to what extent listeners in regular conversational settings perceive breathing.

### 3.4. *Gaze*

In face-to-face interaction, eye gaze has been found to serve many important communicative functions, including referring to objects, expressing intimacy, dominance, and embarrassment, as well as regulating proximity and turn-taking (Argyle and Cook, 1976). It is of course important to note that eye gaze cannot be understood purely as a signaling device, without also considering its main function: to direct our visual attention and perceive the world around us. However, as humans, we have learned through evolution that the eye gaze of others (and by extension their attention) provides a rich source of information for us to coordinate our activities with each other (Tomasello et al., 2007). When two or more people engage in interaction, they typically position themselves in a way that facilitates the monitoring of each others' gaze direction (Schneider and Goffman, 1964). It is therefore possible that we have also learned that eye gaze (to some extent) can be directed to achieve a certain communicative effect.

One of the first extensive analyses of the role of eye gaze in turn-taking was done by Kendon (1967), through observations of video recordings of dyadic interactions. A general pattern he observed is that the speaker tends to look away at the beginning of the turn, but then shift the gaze towards the listener at the end of the turn. At the same time, the listener looks at the speaker for the most part of the turn, but looks away as the turn is being completed and as the turn shifts. If the current speaker pauses without yielding the turn, she is likely to keep looking away, but then look back as the utterance is being resumed. If the listener starts to look away as the speaker's turn is being completed, it can serve as a turn-initial cue, i.e., that she is assuming that the turn is soon yielded and that she is preparing to take the turn. Another finding was that the listener tends to look at the speaker most of the time, and only looks away briefly, whereas the speaker shifts the gaze between the listener and to some other target with about equal lengths. Thus, the listener in general looks more at the speaker than the other way around. Mutual gaze between the participants is seldom maintained for longer than a second.

Similar findings have been reported in several studies (Goodwin, 1981; Oertel et al., 2012a; Jokinen et al., 2010) However, it is also clear that even if these are general patterns, there are other perceptual, communicative and social factor involved, which means that there is a lot of variation in them. There are also large individual differences (Cummins, 2012).

Bavelas et al. (2002) also examined gaze behaviour around backchannels in dyadic interactions, where one person was telling a story to the other. In general, their analysis confirmed Kendon's finding that the listener looks more at the speaker than the reverse, but they also found that the speaker looked at the listener at key points during their turn to seek a response. At these points, the listener was very likely to respond with a verbal or non-verbal backchannel, after which the speaker quickly looked away and continued speaking.

As discussed in Section 2.4 above, gaze also serves an important role for addressee selection as well as next-speaker selection in multi-party interaction (Auer, 2018; Jokinen et al., 2013; Ishii et al., 2016). In this way, gaze towards a participant serves both as a turn-yielding cue and as a signal for selecting the next speaker in multi-party interaction. When addressing several people, the current speaker may alternatingly look at the co-participants they want to address, but then ends the turn by looking at the selected next speaker (ibid.). While people do not always conform to this selection, the gazed-at participant will take the turn more often than not. If the targeted person does not respond, a sustained gaze at that person is particularly efficient to elicit a response (ibid.). If the targeted person wants to avoid taking the turn, they can either pass on the turn by gazing at a third participant, or they can reject the offered turn by gazing away and thereby open up the conversational floor for any other participant to self-select (Weiss, 2018). Zima et al. (2019) investigated the role of gaze for competitive overlap resolution in triadic interactions. They found that when two speakers started to speak at the same time, the prevailing speaker averted their gaze away from the competing speaker. The speaker who withdrew from the competition instead maintained their gaze at the prevailing speaker. The third party often singled out the prevailing speaker during the overlap by either keep looking or shifting the gaze towards her.

In situated interaction, there might also be objects that the speakers refer to, especially in task-oriented settings. When referring to objects, speakers naturally attend to them. The speaker's gaze can therefore be used by the listener as a cue to the speaker's current focus of attention, so-called joint attention (Velichkovsky, 1995). This has been shown to clearly affect the extent to which humans otherwise gaze at each other when speaking and shifting turns (Argyle and Graham, 1976). In a study on modelling turn-taking in three-party poster conversations, it was found that the participants almost always looked at the shared poster and very little at each other (Kawahara et al., 2012). In a Wizard-of-Oz study on multi-party human-robot interaction, where the participants discussed objects on a table between them, Johansson et al. (2013) found that turn shifts often occurred without the speakers looking at each other. This might of course affect the usefulness of gaze as a turn-taking cue in such settings.

### 3.5. Gestures

In the analysis of turn-taking cues by Duncan (1972), certain gestures seemed to have a very strong turn-holding effect (and even negate other turn-yielding cues). The listener almost never attempted to take the turn while the speaker performed certain forms of gesticulation, including a tense hand position or hand movements away from the body. The observation that the completion of hand gestures can serve as a turn-yielding cue has also been confirmed in other studies (Zellers et al., 2016). Holler et al. (2018) investigated how bodily signals influence language processing in interaction. They found that questions accompanied by gestures resulted in faster response times compared to questions without gestures. The response timing also seemed to be oriented with respect to the termination of the gesture. Thus, it seems like gestures help the listener to predict turn-endings. Sikveland and Ogden (2012) also found speakers to temporarily freeze and hold their gesture when some other participant provided some kind of mid-turn clarification or feedback, after which their turn and gesture were resumed.

In the analysis of Streeck and Hartge (1992), gestures by the listener can be used as an indication that they want to get the floor ("early turn-incursion while avoiding overlap"), and as a turn-initial cue. Mondada (2007) investigated turn-taking in a multi-party, situated interaction, involving objects (such as a map) on a shared table. In this setting, pointing gestures (with a finger or a pen), or stretching, towards these objects were also found to serve as a turn-initial cue, signalling an interest in self-selecting the turn. This movement (which typically involves the whole upper body) is often initiated before the completion of the preceding turn, and can thus be used for projection by the other participants.

### 3.6. Summary

As this review has shown, turn-taking cues across different modalities can be both redundant and complementary. The combination of several cues can lead to more accurate recognition or prediction of the partner's intentions, which might help to explain why many people prefer face-to-face interaction. Especially for conversational systems, where the recognition of these subtle cues is challenging, the combination of different cues may increase robustness. As discussed before, this is an argument for why social robots might offer better interaction opportunities than voice assistants (Skantze, 2016).

Verbal cues are arguably the most important ones for humans, and provide a stronger basis for prediction, especially when taking the larger dialogue context into account. However, they are also very hard to model, especially in conversational systems, partly due to the error-proneness and delay of ASR output, and partly due to lack of more sophisticated pragmatic modelling. Thus, prosodic cues can complement verbal cues and help when there are ambiguities. Gaze can be a strong cue, especially in multi-party interaction. However, this requires some form of embodiment that makes it natural to look at the agent in the same way we look at each other during the conversation. Also, if the conversation involves objects in the surroundings, this might significantly decrease the extent to which we look at each other to regulate turn-taking. The use of gestures and breathing for turn-taking in conversational systems and human-robot interaction have so far attracted less attention, and should be worthy to explore.

In the next section, we will discuss attempts at utilising these cues for knowing when the system should speak and not.

## 4. End-of-turn detection and prediction

The arguably most studied aspect of turn-taking in conversational systems is how to determine when the user's turn is yielded and the system can start to speak (i.e. the detection of TRPs). A related aspect is to determine when the system should give a backchannel (i.e., the detection of BRPs, as discussed in Section 2.3). In this section, we will review attempts at developing such models, and also include work done on allowing the system to project turn completion, perform sentence completion, or other forms of overlapping speech, even if these are not very common yet. The use of more explicit cues, such as push-to-talk, will not be covered in this review.

We have divided these approaches into three types (ordered from simpler to more advanced), illustrated in Fig. 4:

- **Silence-based models**. The end of the user's utterance is detected using a VAD. A silence duration threshold is used to determine when to take the turn.
- **IPU-based models**. Potential turn-taking points (IPUs) are detecting using a VAD. Turn-taking cues in the user's speech are processed to determine whether the turn is yielded or not (potentially also considering the length of the pause).
- **Continuous models**. The user's speech is processed continuously to find suitable places to take the turn, but also for identifying backchannel relevant places (BRP), or for making projections.

These models are sometimes referred to as **end-of-turn detection** models (or simply "endpointing") and sometimes as **end-of-turn prediction** models. De Kok and Heylen (2009) argue that the former term should be associated with a more "reactive" account of turn-taking and the latter with a more "predictive" account. Furthermore, both Schlangen (2006) and De Kok and Heylen (2009) argue that silence-based models should be associated with a reactive account, whereas models that take turn-taking

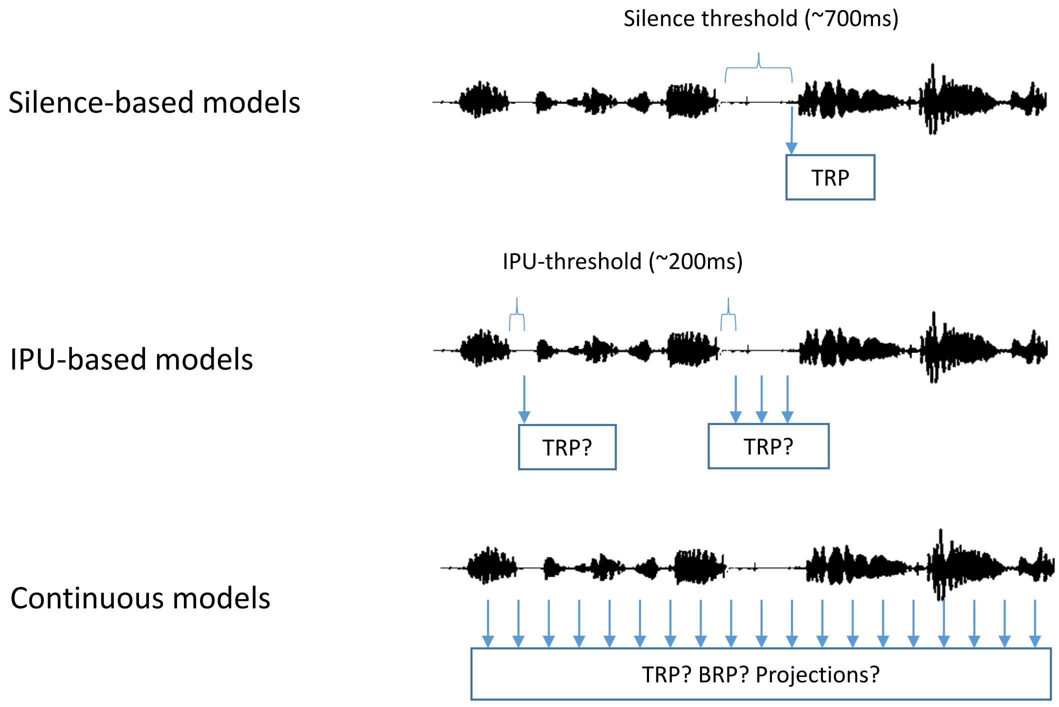


**Fig. 4.** End-of-turn detection and prediction models.

cues into account should be considered predictive. We think this distinction is somewhat misleading and would like to reserve the term "prediction" for models that are associated with predictive mechanisms and projections in turn-taking (which was discussed in Section 2.2). We argue that both IPU-based models and silence-based models can be considered reactive, in the sense that they react to past cues (whether silence-based or more sophisticated) in order to make a decision for the current point-in-time. We will thus use the term *end-of-turn detection* to refer to such models and the term *end-of-turn prediction* for models that predict an upcoming turn-completion (i.e. that has not occurred yet), which is perhaps only feasible with continuous models.

### *4.1. Silence-based models*

In most ASR systems, an end silence duration threshold is often used to determine the end of the speech segment which is to be transformed to text, often relying on a VAD. The VAD is in turn typically based on energy and possibly spectral features (to distinguish speech from background noise) in the incoming audio. Thus, when implementing a conversational system, it is convenient to use this mechanism for determining the end of the user's turn. This approach is still being used in many conversational systems, and was for example assumed to be used in the VoiceXML standard, developed by W3C (McGlashan et al., 2004). Typically, two main parameters (thresholds) can be tuned by the application developer, that regulates the turn-taking behaviour of the system, as illustrated in Fig. 5 (a):

- After the system has yielded the turn, it awaits a user response, allowing for a certain silence (a gap). If this silence exceeds the no-input-timeout threshold (such as 5 s), the system should continue speaking, for example by repeating the last question.
- Once the user has started to speak, the end-silence-timeout (such as 700ms) marks the end of the turn. As the figure shows, this allows for brief pauses (shorter than the end-silence-timeout) within the user's speech.

This basic model might work well when the user's turns are expected to be fairly brief and when the user knows what to say to the system. However, it often results in two kinds of problems: If the end-silence-timeout is too short, the system might interrupt the user within pauses (as illustrated in Fig. 5 (b)). If it is too long, the system will be perceived as unresponsive, and the user might not understand that the system is about to respond (as illustrated in example (c)). To minimise these problems, it is important to carefully tune the end-silence-timeout threshold, depending on the domain of the system (Witt, 2015). It is also possible to use different thresholds depending on the preceding system utterance or dialogue state. For example, a yes/no question can have a shorter threshold than an open question, as the user is more likely to answer with something brief that does not contain pauses. However, as discussed earlier, studies on human-human turn-taking have found that pauses are on average

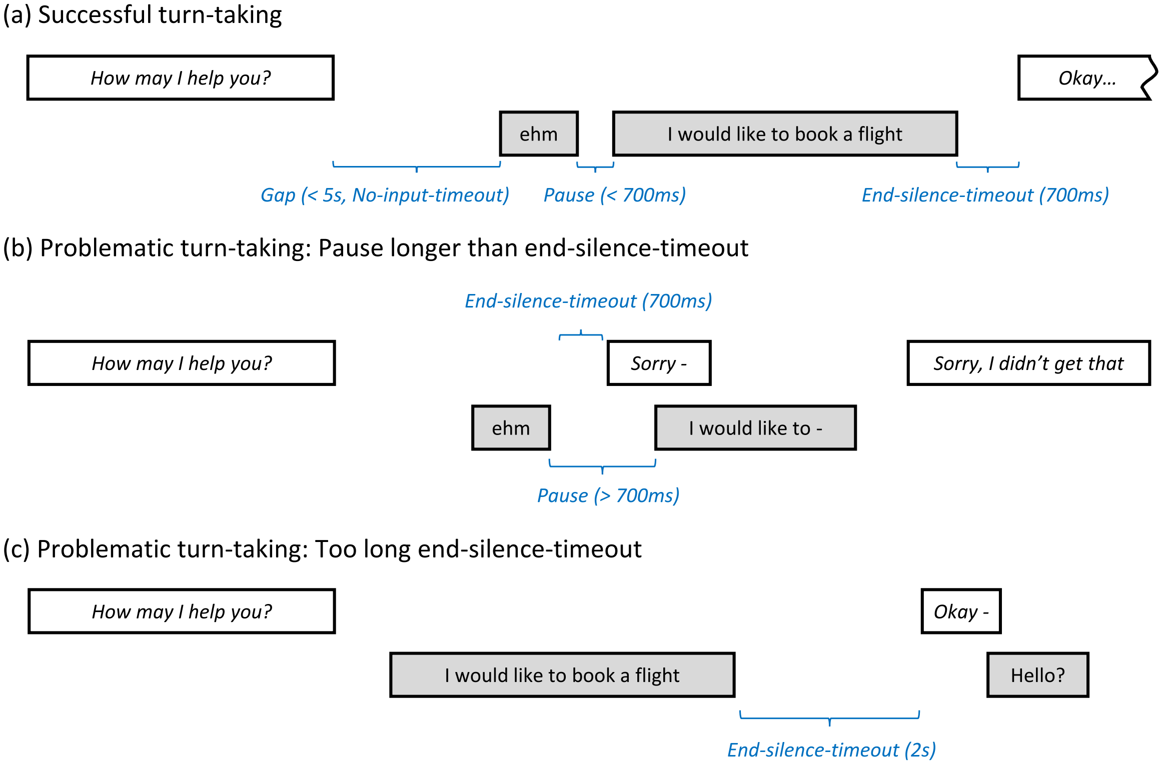

(a) Successful turn-taking
How may I help you?
Okay...
ehm
I would like to book a flight
Gap (< 5s, No-input-timeout)
Pause (< 700ms)
End-silence-timeout (700ms)
(b) Problematic turn-taking: Pause longer than end-silence-timeout
End-silence-timeout (700ms)
How may I help you?
Sorry -
Sorry, I didn't get that
ehm
I would like to -
Pause (> 700ms)
(c) Problematic turn-taking: Too long end-silence-timeout
How may I help you?
Okay -
I would like to book a flight
Hello?
End-silence-timeout (2s)

**Fig. 5.** Silence-based model of turn-taking.

longer than gaps. Thus, it is most often impossible to find a perfect threshold that completely alleviates these two problems, and systems based on this simplistic model will be plagued by a certain amount of turn-taking issues (Ward et al., 2005; Raux et al., 2006).

### 4.2. IPU-based models

Given the limitations of pure silence-based models described above, several researchers have investigated how turn-taking cues could be incorporated in the end-of-turn detection. Based on the assumption that the system should not start to speak while the user is speaking, a common approach has been to detect the end of IPUs in the user's speech using a VAD (in this way similar to the silence-based model outlined above, but potentially using a much shorter silence threshold, such as 200ms). After the end of an IPU has been detected, the system uses the turn-taking cues detected from the user to determine whether there is a TRP or not, as illustrated in Fig. 4. If these TRPs are correctly identified, the system can take the turn with very small gaps, while avoiding to interrupt the user at non-TRPs.

An early example of an IPU-based model was Bell et al. (2001), who used a rule-based approach (a semantic parser) for determining semantic completion after an end-of-speech was detected and the ASR reported a result. If the utterance was determined to be incomplete, the system refrained to take the turn and the ASR continued to listen for more speech. A more data-driven approach, also taking other cues into account, was proposed by Sato et al. (2002), who used a decision tree to classify pauses longer than 750 ms, based on features from semantics, syntax, dialogue state, and prosody. Their model achieved an accuracy of 83.9%, compared to the baseline of 76.2%. Similar models have been explored by Schlangen (2006) and Meena et al. (2014), using shorter silence thresholds.

A problem with using a fixed silence threshold in this approach, after which the decision has to be made, is that it is unclear what will happen if the IPU is misclassified as a pause. Even if a TRP is not detected immediately after an IPU, the system should continue to consider whether there is a TRP as the silence progresses. Intuitively, the longer the silence, the more likely it is that the user is indeed yielding the turn. Ferrer et al. (2002) trained a decision tree classifier based on prosodic features (pitch and duration), as well as n-grams of the final words, but also conditioned their model on the length of the pause after the IPU. The model was then applied continuously throughout the pause, with an increasing likelihood of a turn-shift (as illustrated in Fig. 4). Raux and Eskenazi (2008) formulated the problem somewhat differently. Instead of conditioning the model on pause length, the system used a model to predict a threshold to be used for determining when to take the turn after an IPU was detected. Thus, in case turn-holding cues were detected, more time would be allowed for the user to start speaking again, whereas if turn-yielding cues were detected, only a brief (or no) silence passed before the system took the turn. A similar approach, using deep learning (LSTM), was proposed by Maier et al. (2017).

A general question when training IPU-based models is what target labels to use for detecting TRPs. When using a human-human dialogue corpus, it is of course possible to use the actual turn shifts as target labels. It is however important to note that humans do not always take the turn at TRPs, and sometimes at non-TRPs. Also, models trained on human-human dialogue might not necessarily transfer so easily to human-computer dialogue, as they are a bit different in nature. Using human-computer dialogue data is also problematic if the system used to collect the data employed a more simplistic turn-taking model (as we would not want the new model to learn from this). One approach to solve this is to use bootstrapping. First, a more simplistic model of turn-taking is implemented in a system and interactions are recorded. Then, the data is manually annotated with suitable TRPs, and a machine learning model is trained (cf. Raux and Eskenazi 2008; Meena et al. 2014; Johansson and Skantze 2015). Another approach is to use a Wizard-of-Oz setup, where a hidden operator controls the system and makes the turn-taking decisions (Johansson et al., 2016; Maier et al., 2017). As the Wizard is expected to take the turn at appropriate places, the data can be used directly to train a model. A potential problem with this approach is that it can be hard for the Wizard to achieve very precise timing.

Another approach is to use reinforcement learning. Jonsdottir et al. (2008) showed how two artificial agents could develop turn-taking skills by talking to each other, learning to pick up each others' prosodic cues. In the beginning, pause durations were too short and overlaps were frequent. However, after some time, interruptions were less frequent and the general turn-taking patterns started to resemble those of humans. Selfridge and Heeman (2010) presented an approach where turn-taking was modelled as a negotiative process. In this model, each participants "bids" for the turn, based on the importance of the intended utterance, and reinforcement learning was used to indirectly learn the parameters of the model. Khouzaimi et al. (2015) used reinforcement learning to learn a turn-taking management model in a simulated environment, with the objective of minimising the dialogue duration and maximising the completion task ratio. However, it is perhaps not clear to what extent such a model transfers to interactions with real users. Ideally, the system should be able to learn to detect the end of the turn through interaction with users.

Looking at the cues that have been found to be useful for turn-taking detection in the context of a dialogue system, different studies have come to different conclusions, likely depending on the domain of the system and how well these turn-taking cues are recognised and modelled. Whereas Sato et al. (2002) and Meena et al. (2014) did not find prosody to contribute significantly to the detection, Ferrer et al. (2002) and Schlangen (2006) found both syntactic and prosodic features to improve detection performance. As we saw in Sections 3.4 and 3.5, face-to-face interaction also allows for visual turn-taking cues, which were used in the model presented in De Kok and Heylen (2009). Johansson and Skantze (2015) built an IPU-based model for multi-party human-robot interaction, and investigated the use of words, prosody, head pose (as a proxy for gaze), dialogue context, and the movement of objects on the table. All these cues were in themselves informative, but the best performance was achieved by combining them, which is in line with findings in the linguistic literature.

A challenge when using verbal features in dialogue systems is speech recognition (ASR) errors. This is especially troublesome when considering the syntactic or pragmatic completion of the final part of the utterance, since ASR language models are typically trained on syntactically complete sentences, and therefore might miss important turn-holding cues at the end of the utterance, such as a preposition ("I would like to go to...") or a filled pause. Some speech recognition vendors (like Google at the time of writing this review) do not even report filled pauses (which are typically very strong turn-holding cues) in the user's speech. Meena et al. (2014) explored the effect of ASR errors on their IPU-based model. Whereas prosody had very little contribution to the performance when the ASR was perfect, it did contribute when ASR performance degraded.

### *4.3. Continuous models*

Most conventional dialogue systems (either silence- or IPU-based) typically process the user's speech one utterance (or one IPU) at a time. A VAD is used to detect the end of the utterance, after which it is processed, one module at a time, and a system response is generated (unless the system refrains from taking the turn), as illustrated in the left pane in Fig. 6. In addition to the delay caused by the silence threshold, the processing time of each subsequent module adds to the response delay. An alternative approach is to process the dialogue incrementally, as shown in the right pane. This means that the input from the user is processed in increments (e.g. frame-by-frame or word-by-word), and that all modules start to process the input as soon as possible, passing on their incremental results to subsequent modules (Schlangen and Skantze, 2009).

Incremental processing allows the system to process the user's utterance on a deeper level (also incorporating task-related aspects), and make continuous turn-taking decisions. Thus, the system can potentially also project turn-completions, starting planning what to say next, find suitable places for overlapping backchannels, and even decide to interrupt the user. All this is impossible with an IPU-based model. An early example of a fully incremental dialogue system was presented by Skantze and Schlangen (2009), although it was constrained to number dictation. While the user was reading out the numbers, the system could start to prepare responses and give very rapid feedback, based on continuous processing of the user's speech (including prosodic turn-taking cues). One important challenge of incremental systems identified by Schlangen and Skantze (2009) is the need for *revision*, as tentative hypotheses of what the user is saying might change as more speech is being processed. For example, the word "four" might be amended with more speech, resulting in a revision to the word "forty". If subsequent modules have already started to make output results based on the earlier hypotheses, they might also have to make revisions in their output, potentially causing a cascade of revisions. Ultimately, if the system has already started to speak when a revision occurs, it might have to synthesize self-repairs, as explored by Skantze and Hjalmarsson (2013).

Skantze (2017b) proposed a general continuous turn-taking model, which was trained on human-human dialogue data in a self-supervised fashion. The audio from both speakers was processed frame-by-frame (20 frames per second), and an LSTM was trained to predict the speech activity for the two speakers for each frame in a future 3 s window. Thus, the model was not trained specifically for end-of-turn detection. However, when applied to this task, the model performed better than more conventional baselines, and better than human judges given the same task. The model was also able to make predictions about utterance length at the onset of the utterance, which could potentially be useful for distinguishing attempts to take the turn from shorter backchannels. A similar model was implemented by Ward et al. (2018) and Roddy et al. (2018a), who also looked more deeply into the different speech features that can help the prediction. Roddy et al. (2018b) proposed an extension of the architecture, where the acoustic and linguistic features were processed in separate LSTM subsystems with different timescales.

As will be discussed in Section 7.3, models based on incremental processing are also highly relevant for interactions involving some form of task execution, for example where a human gives instructions to a robot, where the execution of the task can be initiated already while the instruction is being given (Hough and Schlangen, 2016; Gervits et al., 2020). Another possibility for more sophisticated turn-taking that incremental processing allows for was shown by DeVault et al. (2009), where the system predicted what the user was about to say in order to help the user to complete the sentence, possibly overlapping with the user's speech.

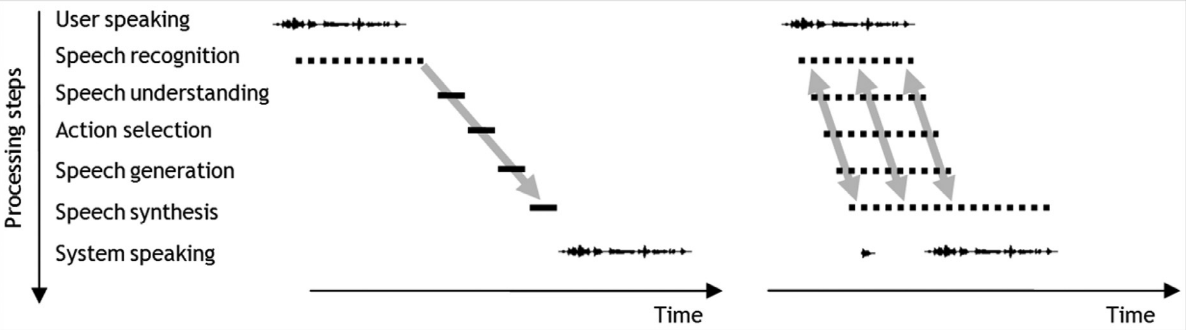


**Fig. 6.** Illustration of the difference between traditional non-incremental speech processing (on the left) and incremental speech processing (on the right).

### 4.3.1. *Detecting backchannel-relevant places*

Apart from detecting appropriate places to take the turn, the system may also consider where to produce backchannels, i.e. detecting BRPs. As discussed in Sections 2.3 and 3 above, the cues that indicate backchannel-relevant places are somewhat different from turn-yielding cues. It should also be noted that backchannels can be non-vocal, such as head nods. While it is possible to use an IPU-based model for backchannels (cf. Truong et al. 2010), a continuous model is more appropriate, as backchannels often are produced in overlap while the other participant is speaking.

Most backchannel models have been based on non-verbal cues, which avoids the problems of dealing with ASR in continuous models. As mentioned in Section 3.2, Ward (1996) defined a very simple rule for backchannel-relevant places: 200ms after a region of low pitch (defined as less than the 30th percentile pitch level), allowing for overlapping backchannels. The model was implemented in a semi-autonomous dialogue system and informally evaluated. A more probabilistic approach was proposed by Morency et al. (2008), where a sequential probabilistic model was trained on human-human face-to-face interactions to predict listener backchannels (in the form of head nods), using prosody, words and eye gaze. The model is continuous, in that it outputs a probability of a backchannel for every frame. By setting a threshold on this probability, the system can generate backchannels at appropriate places, while the frequency of backchannels can be adjusted. The model was directly compared with the rule-based approach of Ward (1996), and showed a significant improvement.

More recently, deep learning approaches have been proposed. Ruede et al. (2019) used an LSTM to predict (vocal) backchannels in the Switchboard corpus, based on prosodic features as well as linguistic information (in the form of word vectors). Hussain et al. (2019) addressed the problem of backchannel-generation (including laughter) in the context of human-robot interaction. Instead of a supervised approach, they used deep reinforcement learning to maximise the engagement and attention of the user.

## 5. Handling user interruptions and backchannels

Although dialogue systems can be implemented using a simplex channel (i.e., where only one participant can speak at a time), it is often desirable to have a duplex channel, allowing the system to hear what the user is saying while it is speaking (i.e.

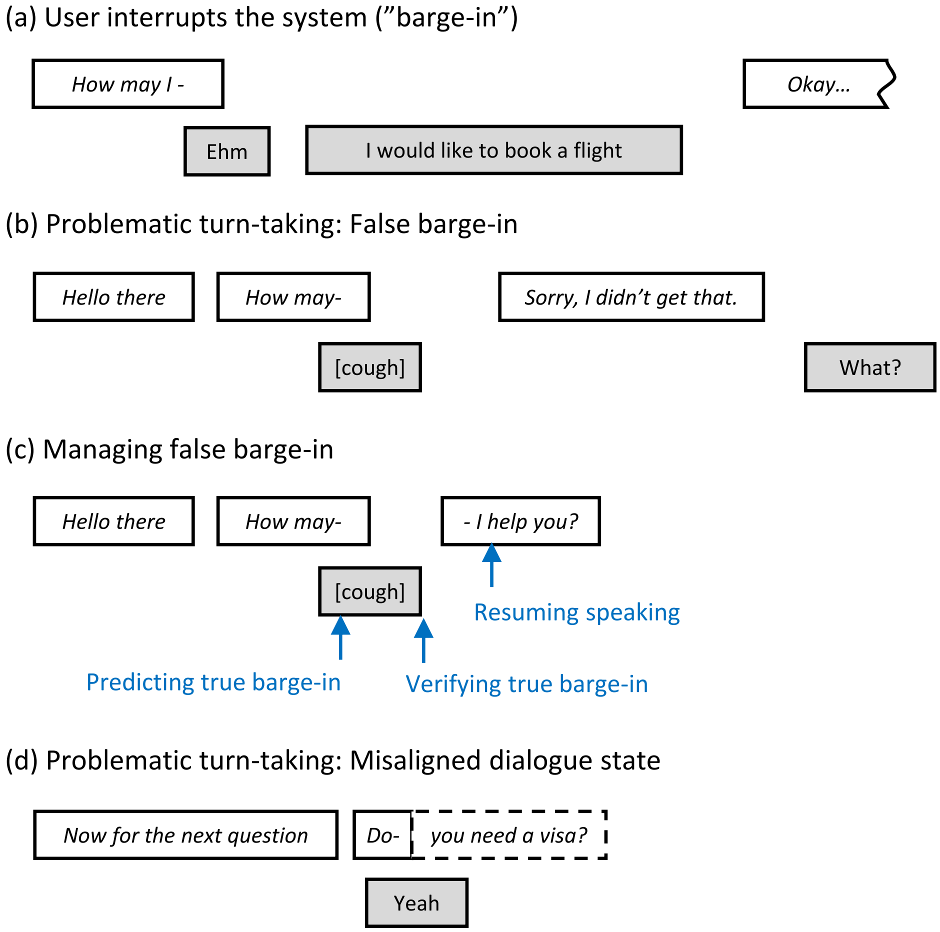

(a) User interrupts the system (”barge-in”)
How may I -
Okay...
Ehm
I would like to book a flight
(b) Problematic turn-taking: False barge-in
Hello there
How may-
Sorry, I didn’t get that.
[cough]
What?
(c) Managing false barge-in
Hello there
How may-
- I help you?
[cough]
Resuming speaking
Predicting true barge-in
Verifying true barge-in
(d) Problematic turn-taking: Misaligned dialogue state
Now for the next question
Do-
you need a visa?
Yeah

**Fig. 7.** The handling of user-interruptions, and associated problems.

overlapping speech). A requirement for this is that the system uses echo cancellation, i.e., that the system's own voice is cancelled out from the audio picked up by the microphone. The most common use case for this is to allow the user to "**barge-in**" while the system is speaking, i.e., to interrupt the system, as illustrated in Fig. 7(a). This is especially important when the system asks longer questions or gives longer instructions which the user might have heard before or can be predicted from context. However, there are also several caveats associated with barge-in.

One of the first to perform studies of users' barge-in behaviour was Heins et al. (1997). In their study, they found that users attempted to barge-in without being informed about the possibility and did so frequently. Barge-in attempts tended to happen at certain places in the system's prompt, especially at syntactic boundaries. They also noticed that user disfluencies (in the form of stuttering or repetitions) were common in barge-in situations, which is in line with the observations of competitive overlap in human communication discussed in Section 2.3. This can of course cause problems for further processing of the user's speech.

A typical problem when allowing barge-in is that of false barge-ins, as illustrated in example Fig. 7(b). A false barge-in might either be triggered by non-speech audio, such as external noise or coughing, or by the user giving a backchannel without intending to take the turn. If not handled correctly, this can easily lead to confusion. One way of reducing this problem suggested by Heins et al. (1997) is to raise the threshold for detecting user speech when this is less likely to occur and lower it when user interruptions are more likely (for example at a syntactic boundary). Apart from noise and coughing, the user might also just produce a brief backchannel, without intending to take the turn. The system should therefore as early as possible try to predict, already at IPU onset, whether the incoming audio is likely to constitute a longer turn or not. If not, it probably does not even have to stop speaking. One example of such a model (trained on human-human data) was presented by Neiberg and Truong (2011), where a prediction of a turn-start vs. a backchannel was made 100 − 500 ms after the speech onset, based on acoustic features. Another example is the general model of Skantze (2017b), which was also shown to be able to make such distinctions by predicting the length of the user's utterance.

If the user's utterance was determined to be a true barge-in, it should still be verified at the end of the IPU. If it was not a true barge-in, the system should ideally resume speaking, as illustrated in Fig. 7(c). As suggested by Ström and Seneff (2000), this could for example be done using a filled pause, and then restart from the last phrase boundary. An example of how these different aspects are integrated is presented in Selfridge et al. (2013), where incremental speech recognition was used to make a continuous decision of whether to pause, continue, or resume the system's utterance. The model was evaluated in a public spoken dialogue system, resulting in improved task success and efficiency compared to a more naive approach.

Another potential problem when allowing for barge-in is illustrated in Fig. 7(d). Here, the user responds to the first system utterance ("Now for the next question") with "Yeah", before hearing that the system has just started the next utterance ("Do you need a visa?"). Depending on how the dialogue manager is implemented, there is a risk that the user's "Yeah" will be interpreted as a positive reply to the question, which will be very confusing as the user has never heard the question which was unintentionally answered. To mitigate this problem, the system should ideally monitor its own speech production in order to evaluate the likely context in which to interpret the user's utterance.

Heins et al. (1997) also noted that users sometimes barge-in before hearing the important things that they need to hear. Thus, it is also possible that the system should not always allow for barge-in, for example if it has something important to say. Ström and Seneff (2000) explored how this could be signaled to the user when a barge-in attempt is detected. By raising the volume of the system's voice (in line with human behaviour during competitive overlap discussed in Section 2.3), it can signal that the barge-in is not allowed, and correspondingly lower the voice if it is allowed.

## 6. Generating turn-taking cues

So far, we have mainly reviewed the processing of turn-taking cues from the user. However, it is also important to consider the generation of turn-taking cues, so that the user knows when it is appropriate to speak and not. If the system fails to do this correctly, the user might start to speak at the same time the system. An example of an unclear turn-allocation is illustrated in Fig. 8(a). If the system makes a pause (for example at a phrase boundary, or because it needs time for processing), it is important that an appropriate turn-holding cues is produced. In the example in Fig. 8(a), an appropriate prosodic realisation (using level intonation as discussed in Section 3.2) might be sufficient. However, conversational systems often do not have this level of control of the speech synthesizer, which might cause problems. Thus, even if turn-taking cues are not generated on purpose, users will likely interpret these cues from the system's synthesized speech and other behaviours, such as gaze in case of an avatar or robot. Several studies have therefore looked into how to generate appropriate behaviours in animated agents to facilitate turn-taking (Cassell et al., 2001; Thórisson, 1999; Pelachaud et al., 1996).

To explore the effect of such turn-taking cues, Edlund and Beskow (2009) set up an experiment where two participants were connected remotely to have a conversation, with an avatar (an animated head) representing each person on the other end. The audio was transmitted between the participants with lip-sync added automatically to the avatar. The gaze direction of the avatar, however, did not reflect the other participant, but was manipulated for the sake of the experiment. The results showed that when the avatar gazed away at a potential turn-ending, the other participant was less likely to take the turn, compared to when the agent looked at the other participant. Kunc et al. (2013) explored the effectiveness of visual and vocal turn-yielding cues in dialogue system using an animated agent. Their results indicated that the visual cues were more effective than vocal cues. In another study, Skantze et al. (2014) investigated the effect of gaze, syntax and filled pauses as turn-holding cues at pauses in a human-robot interaction setting, where the robot was instructing the user to draw a route on a map that was placed on the table between them. When the robot looked at the map (i.e. did not gaze at the user), used a syntactically incomplete phrase, or used a

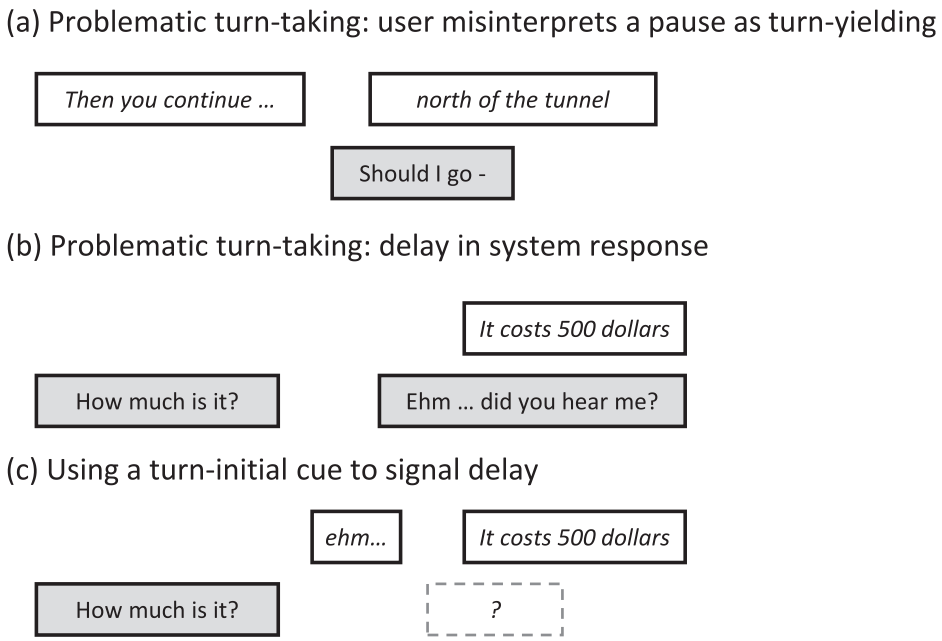

(a) Problematic turn-taking: user misinterprets a pause as turn-yielding
Then you continue ...
north of the tunnel
Should I go -
(b) Problematic turn-taking: delay in system response
It costs 500 dollars
How much is it?
Ehm ... did you hear me?
(c) Using a turn-initial cue to signal delay
ehm...
It costs 500 dollars
How much is it?
?

**Fig. 8.** Problems with lack of turn-taking cues from the system.

filled pause, the user was less likely to continue drawing or give feedback than when the robot looked at the user or produced a complete phrase.

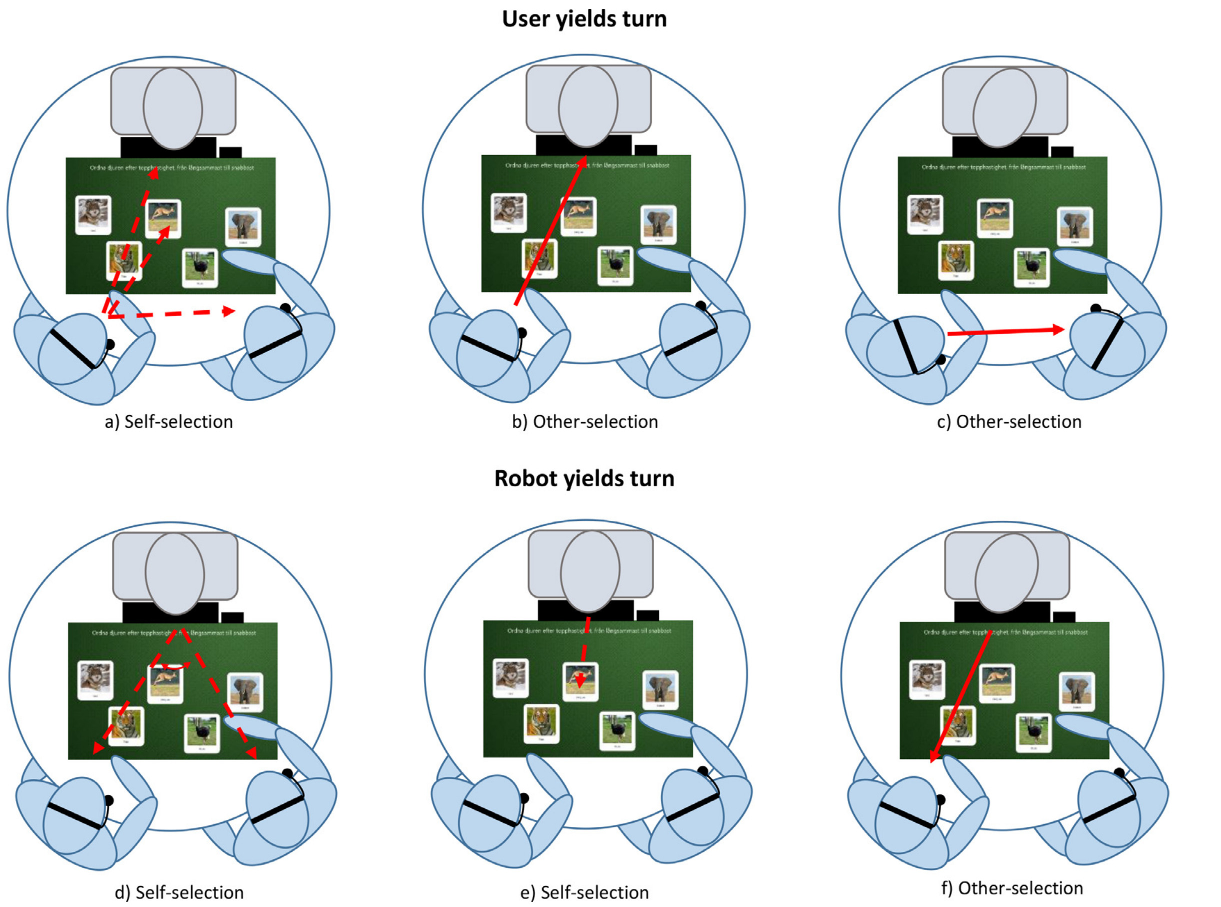
Another potentially problematic case is if there is a delay in the system's processing, which causes a delay in its response to the user. In such cases, the user might not understand that the system is about to speak, and might continue speaking, as illustrated in Fig. 8(b). As discussed in Sections 2.1 and 2.3 above, humans typically use various turn-initial cues in these situations – such as a filled pause, in-breath, gaze aversion, to signal their willingness to take the turn.

Thus, one solution to this problem is to use more shallow (or incremental) processing to find TRPs and decide that the system should respond. At this point, the system can start to produce a turn-initial cue, even if the system does not know exactly what to say yet. Then, when processing of the user's utterance is complete, the system can produce the actual response (cf. Skantze and Hjalmarsson 2013; Skantze et al. 2015; Lala et al. 2019). To investigate the effectiveness of such cues in a human-robot interaction scenario, Skantze et al. (2015) systematically investigated different multi-modal turn-holding cues. Fig. 8(c) shows an example where the user asks a question, and the system is not ready to respond immediately, and where a filled pause is used as a turn-holding cue. To measure the effectiveness of different turn-holding cues, the probability for the user to continue speaking in the window marked with "?" (which should be avoided) can be calculated. It was found that all investigated cues were effective: filled pause, in-breath, smile, and gaze aversion. The strongest effect was achieved by combining several cues; gaze aversion together with a filled pause reduced the probability of the user starting to speak by half, compared to using no cues at all. This indicates that the subtle cues humans use for coordinating turn-taking can be transferred to a human-like robot and have similar effects, without any explicit instructions to the user.

Hjalmarsson and Oertel (2011) investigated the effect of gaze as a backchannel-inviting cue (as discussed in Section 3.4 above). They designed an experiment where participants were asked to provide feedback while listening to a story-telling virtual agent. While the agent looked away for the most part while speaking, it gazed at the user at certain points. The results showed that listeners were indeed more prone to give backchannels when the agent gazed at them, although there was a large variation in their behaviour, indicating that there are other important factors involved.

## 7. Modelling turn-taking in multi-party and situated interaction

In Section 2.4, we discussed how turn-taking in multi-party interaction differs from dyadic interaction. Even though voice-only interaction can be multi-party, the natural setting for such interaction is face-to-face, preferably physically situated, interaction. Thus, the modelling of multi-party interaction has mostly been studied in the context of human-robot interaction (although multi-agent, single-user interactions have also been modelled in virtual environments; Traum and Rickel 2002). Apart from the detection and generation of turn-holding and turn-yielding cues, which is relevant for any type of spoken interaction, multi-party interaction also involves the identification of the addressee of utterances. This means that the system has to both detect whom a user might be addressing, and display proper behaviours when addressing a specific user, as illustrated in Fig. 9. We will discuss both these issues in this section, as well as turn-taking in interactions that involve physical manipulation of objects.



User yields turn

a) Self-selection

b) Other-selection

c) Other-selection

Robot yields turn

d) Self-selection

e) Self-selection

f) Other-selection

**Fig. 9.** Possible turn-transitions in multi-party human-robot interaction.

### *7.1. Addressee detection*

Several studies have looked into how to combine different cues for addressee detection in multi-party interaction (i.e., detecting the addressee of a user's utterance), using machine learning (Katzenmaier et al., 2004; Vinyals et al., 2012; Jovanovic et al., 2006). The most obvious signal is clearly the user's gaze. However, for practical systems, tracking the eye gaze of users is non-trivial. Eye tracking equipment typically requires some form of calibration, they are often limited in terms of field-of-view, and are sensitive to blinking and occlusion. Many systems therefore rely on head pose tracking as a proxy for gaze direction, which is a simpler and more robust approach, but which cannot capture quick glances or track more precise gaze targets. Despite this, studies have found head pose to be a fairly reliable indicator of visual focus of attention in multi-party interaction, given that the targets are clearly separated (Katzenmaier et al., 2004; Stiefelhagen et al., 2002; Ba and Odobez, 2009; Johansson et al., 2013).

In addition to gaze, some studies have also found other modalities to contribute. In multi-party interaction (involving two humans and a robot), Shriberg et al. (2012) found that speech addressed towards the machine is louder and more clearly articulated, which can be utilised to make the distinction. This is also in line with the findings of Katzenmaier et al. (2004), although they found visual cues to be more informative.

Johansson and Skantze (2015) explored how the problem of addressee detection could be combined with end-of-turn-detection. They used an IPU-based model (see Section 4.2 above), and annotated each IPU in their dataset (a conversational game between one robot and two humans) according to how appropriate it was for the robot to take the turn. Instead of a binary notion, they used a scale ranging from “not appropriate” to “obliged”. Cases where it is not appropriate to take the turn could either be because the current speaker did not yield the turn, or because the turn was yielded to the other human (Fig. 9c). The robot could also be “obliged” to take the turn, for example if a user looked at the robot and asked a direct question (Fig. 9b). In between these, there were cases where it is possible to take the turn “if needed”, and cases where it is appropriate to take the turn, but not obligatory. These were often cases where the user was attending both the robot and the other user, or objects on the table (Fig. 9a). Head pose (as a proxy for gaze) was a fairly informative feature, 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. By adding more features, such as prosody and verbal features, the performance was improved further.

### *7.2. Addressing users and regulating the interaction*

When it comes to generating visual turn-taking cues, an animated face might be sufficient for dyadic interaction, as the main turn-regulatory function of gaze in such settings is to signal whether the agent is yielding or holding the turn (i.e., looking towards the user or looking away). However, in multi-party interaction, an animated agent on a flat screen might not suffice, as the users and the agent are not sharing the same physical space. This makes it impossible for the user to see exactly where the agent is looking (in the users' physical space), a problem typically referred to as the "Mona Lisa effect" (Al Moubayed et al., 2012). Thus, in a multi-party setting, this means that the agent cannot establish exclusive mutual gaze with one of the users, and in a situated interaction the object that is attended to cannot be inferred. Al Moubayed and Skantze (2011) compared an animated agent on a 2D display with the Furhat robot head, which uses back-projection on a 3D mask, in a multi-party setting to explore the impact on turn-taking. It was found that the turn-taking accuracy (i.e. how often the addressed user was the one taking the turn) was 84% in the 3D condition, vs. 53% in the 2D condition. The response time was also faster in the 3D condition (1.38s vs. 1.85s), which indicates that there was less confusion regarding the gaze target. It has also been shown that humans can utilise the robot's gaze to disambiguate references to objects to achieve joint attention (Skantze et al., 2014).

Several studies have found that robots in multi-party interaction can influence the turn-taking by actively selecting the next speaker using gaze, which the users typically conform to Mutlu et al. (2012), Bohus and Horvitz (2010) and Skantze et al. (2015). This can be used to regulate the interaction, for example to increase the participation equality. One example of this is Andrist et al. (2013), where a virtual agent was used to balance the contributions from a group of children playing a game. Such balancing also requires a measure of dominance or participation equality. In Strohkorb et al. (2015), a model for classifying children's social dominance in group interactions is presented. Based on manual annotation of social dominance, a model was trained. The main predictor that turned out to be useful was the children's gaze towards the robot (although there was no two-way spoken interaction between the robot and the children). Another example is Nakano et al. (2015), where a dominance estimation model (based on gaze and speech) was used to decide how the robot should regulate the interaction using gaze (in a Wizard-of-Oz setup).

Skantze (2017a) investigated the ability of the robot to shape the interaction in a more open three-party conversation. The results showed that the speaking time for most pairs of speakers was fairly imbalanced (with one participant speaking almost twice as much as the other). However, in line with other studies, the robot was able to reduce the overall imbalance by addressing the less dominant speaker (Fig. 9f). This effect was stronger when mutual gaze was established between the robot and the addressee. When the floor was open for self-selection (Fig. 9d-e), the imbalance instead increased.

### *7.3. Turn-taking in a physical world*

Most models of turn-taking only consider speech and other communicative signals, such as gaze and gestures. However, in situated interaction, humans naturally make use of what Clark (2005) refers to as "material signals". This includes the placement of objects (or themselves) in special sites for the addressee to interpret. For example, when a customer in a shop places an item to buy on the counter, this can be interpreted as the dialogue act "I would like to buy this item". Sometimes such actions are accompanied with a spoken utterance (before, during or after the action takes place), but sometimes not. Thus, models of turn-taking in a physical world should also take such actions into account. This might be especially important for human-robot interaction, which often involves physical tasks. For example, Hough and Schlangen (2016) explored a setting where a human gave instructions to a robot arm to move objects on a table, where the robot did not speak at all, but where the actions of the robot provided feedback on what the robot had understood, and where the timing of the instructions and the movements was important for the fluidity of the interaction. Furthermore, the execution of physical actions can be done in overlap with the spoken instruction, and might need some time for preparation, and thus the need for incremental processing and predictive mechanisms becomes more important, as the robot can start to prepare and execute the action before the instruction is complete (Gervits et al., 2020).

Another potential settings for human-robot interaction is that of joint assembly, where only one person at a time might be able to perform an action (Calisgan et al., 2012), and this turn-taking has to be coordinated just like the taking of turns speaking. Calisgan et al. (2012) performed an experiment where humans were given a joint assembly task, in order to identify the signals used to coordinate their turn-taking. Common turn-yielding signals included putting the hands on the table, crossing the arms, or taking a step back. Often, several cues were combined.

A related phenomenon also relevant for human-robot interaction is that of hand-overs. Moon et al. (2014) did a comparative study on how the robot's gaze behaviour affects the efficiency with which the robot can hand over an object to a human. It was found that the most human-like behaviour – where the robot first gazes at the location where the hand-over will take place, and then up at the human when the human reaches for the object – was more efficient. In this condition, the human moved the hand to the hand-over location even before the robot's hand reached there.

In case the task execution is accompanied with a spoken utterance (such as an acknowledgement), the spoken utterance may provide information on the timing of the task execution. Skantze et al. (2014) explored the setting of a robot giving giving instructions to a human subject on how to draw a route on a map. Each piece of instruction was typically followed by an acknowledgement from the human. However, the lexical choice ("okay", "yeah", "mhm", etc.) and prosody of that acknowledgement varied depending on whether it was produced before, while or after the corresponding route segment was drawn. Thus the form of acknowledgement seemed to serve to help the instruction giver to know the timing of the drawing action, and when the next piece of instruction should follow.

When developing end-of-turn detection models (as discussed in Section 4) in interactions involving physical tasks, it is of course also possible to include the manipulation of objects as a turn-taking cue. An example of this is the end-of-turn detection model of Johansson and Skantze (2015), in the context of a robot playing a card-sorting game with two users, which also incorporated the movement of the cards as a feature for the data-driven turn-taking model.

## 8. Summary and future directions

As this review has shown, turn-taking in dialogue is a highly complex phenomenon that has attracted the interest of researchers from multiple disciplines. The coordination of when a turn will transition, and who will speak next, relies on multiple signals across different modalities, including syntax, pragmatics, prosody and gaze. To some extent these cues are redundant, and to some extent complementary, but the relative merits of different signals still needs to be investigated further. Conversational systems have traditionally relied on simple silence timeouts to identify suitable turn transitions, but this is clearly not sufficient, as pauses within turns are frequent. Several studies have shown how turn-taking cues from the user can be identified to better detect or predict when the turn is yielded. It is also important to consider how the system can make use of such signals to help the user understand when the floor is open and not, and who is supposed to speak next. If the agent is embodied, the visual channel adds more potential cues which can further improve the coordination of turn-taking.

However, despite this progress in computational modelling of turn-taking, it is interesting to note how simplistic turn-taking models in most state-of-the-art conversational systems still are. Recently, there has been a surge in the research on neural models for dialogue systems and chatbots, in both academia and industry, especially regarding dialogue state tracking and end-to-end response generation. The target channels where these models have been deployed and evaluated have typically been either written chats, where turn-taking is trivial, or voice assistants in smart phones and smart speakers, where explicit turn-taking using wake words has become a norm (for various reasons, as discussed in the introduction). This might help to explain why turn-taking has not attracted the same level of interest as these other problems. We think this might change as the applications of spoken dialogue systems move beyond question-answering and command-and-control, and towards more conversational interaction styles, as well as social robots that exhibit a wider repertoire of turn-taking signals and interaction settings.

In the rest of this section, we will discuss some potential directions for future research on turn-taking in conversational systems.

### *8.1. General and adaptive models*

A problem with current models of turn-taking is that they are trained and evaluated on specific (often human-human) corpora, such as Switchboard (Godfrey et al., 1995) or Map Task (Anderson et al., 1991). Thus, there is of course a risk that these models learn the patterns and peculiarities of these corpora, and it is not clear how well they actually generalise to other dialogue styles, and especially to human-computer dialogue. As was discussed in Section 4, models that have been applied to actual dialogue systems have typically been trained on data of interactions with those specific systems. In addition, these models have required some form of labelling, either through manual annotation or through Wizard-of-Oz. This is typically too costly for the development of most dialogue systems, and might help to explain why such models are not widely used in practical systems.

Of course, not only the interaction setting and style affects turn-taking behaviour, but there are also cultural, demographic and individual differences, as discussed in Section 2.5. Thus, it is important that future research investigates how general models can be trained on various types of interactions, and how well such models perform on human-computer interactions of different sorts. One approach would be to use some form of self-supervised (Skantze, 2017b) adaptation of the model, by letting the model make predictions about the future on a certain interaction data set, if that is available. Another approach would be to use some form of reinforcement learning (Jonsdottir et al., 2008; Selfridge and Heeman, 2010; Khouzaimi et al., 2015). Whether such models could be used to fine-tune a turn-taking model in interaction with humans is of course a very interesting question to explore.

### *8.2. Achieving natural timing*

Most models of turn-taking in conversational systems have been based on the assumption that the system should be able to respond as quickly as possible. However, if this goal is achieved, this might not result in very natural conversations. In human-human dialogue, response time naturally depends on factors such as cognitive load, personality and situation (Strömbergsson et al., 2013). While too long response time can lead to uncomfortable silence and turn-taking problems (as discussed in Section 4.1), too quick responses can be perceived as insincere or unreflected. Thus, a certain delay might be more appropriate for certain questions, and this delay should depend on the preceding context and the type of response. Looking at human-human dialogue, it is clear that response time varies depending on the type of dialogue acts exchanged. In an analysis of the Switchboard corpus, Strömbergsson et al. (2013) found that both question type and response type affected the response time, where open and wh-questions had a longer response time than yes/no and alternative questions (300–450ms vs. 100–180ms). Stivers et al. (2009) found that confirmations are delivered 100-500ms faster on average, compared to disconfirmations, in a study of 10 different languages.

A recent model that incorporates this aspect was presented by Roddy and Harte (2020), where a continuous end-of-turn detection model uses an encoding of both the previous utterance (from the user) and the upcoming response (from the system) to incrementally predict response timings, based on human-human dialogue data.

### 8.3. Taking pragmatics and utility into account

As the review on turn-taking cues in Section 3 showed, syntactic and pragmatic completion plays a major role in turn-taking. However, pragmatic completion is very hard to model. Current approaches are often very simplistic, for example relying on the last POS tags in the utterance. Either much stronger language models are needed, which can take the larger dialogue context into account (as discussed in Section 3.1), or the system needs to involve deeper processing of the user's speech, where all components of the dialogue system is involved in the decision of when to take the turn and not. This requires an incremental processing framework, as discussed in Section 4.3.

When training turn-taking models on task-oriented dialogue data, such as Map Task, it is clear that information related to the task is missing. In the example shown in Fig. 1, speaker B responds quickly to the first question ("have you got an extinct volcano?") because speaker B can identify the object on her map. If that would not have been the case, the response time would have been different, and this is not information that is accessible to the turn-taking model when trained purely on speech data.

In general, people do not just take the turn because there is a TRP, but because they have something to say. Thus any turn-taking model for task-oriented dialogue should also take the task into account and involve the utility of speaking. If the utility is high (e.g. "there is a fire!"), the system might want to speak regardless of whether there is a TRP. If the utility is low, it might be less prone. However, even if it does not have something important to say, it might still be obliged to take the turn, given a strong enough TRP (e.g. after someone has asked it "What do you think?"). Decision-theoretic models of turn-taking (which involve a notion of utility) have been proposed by Raux and Eskenazi (2009) and Bohus and Horvitz (2011).

### 8.4. Predictive modelling

As discussed in Section 4.3, most turn-taking models are still reactive, in the sense that they do not try to predict the end of the user's turn before it is actually complete. While reactive models may be sufficient for many scenarios, as conversational systems do not have the same cognitive and physiological constraints as humans, and therefore do not have to prepare their response in advance, there are several potential applications for such models. For example, as discussed in Section 7.3, physical actions of robots do have constraints and may need to be planned for in advance.

Beyond turn-shifts, predictive models can also be trained to predict other types of future events, such as prosody, sentence-completions and dialogue acts. This could for example be used by the system to predict how the user will react to the system's own actions. This way, the system could explore different potential future scenarios and plan ahead accordingly. For example, by simulating different prosodic realisations of an utterance, it could evaluate how the user is likely to react, and choose the most appropriate one.

When it comes to prediction, it is still unclear what types of predictions humans make. The model by Skantze (2017b) was based on prediction of speech activity in a 3 second window. A problem with that approach was that the predictions made towards the end of this window were not very accurate. As argued by Sacks et al. (1974), and others, predictions are likely centered on words, where syntactic and pragmatic completion could help. However, the timing of those words is also important in order to achieve accurate coordination.

Another motivation for predictive modelling is to be able to produce cooperative overlaps. While most models of turn-taking in conversational systems are still based on the goal of minimising gaps and overlaps, cooperative overlaps in human-human interaction do contribute to increased fluency and feeling of rapport, as discussed in Section 2.3. Apart from the generation of backchannels, there is so far very little work on how to produce cooperative overlapping speech. A notable exception was DeVault et al. (2009), who developed a model for predicting what the user was about to say, in order to help the user to complete the sentence, possibly overlapping with the user's speech. However, in their experience when testing the system with users, this often resulted in the agent being perceived as barging in and interrupting the user's speech. Thus, given the current state of the technology, the behaviour was deemed to be undesirable in most cases.

Generating cooperative overlapping speech, like the choral talk discussed in Section 2.3, would likely be an extremely challenging task, as it involves both the incremental processing of the user's speech, prediction of how it is likely to continue, as well as the incremental generation of synthesized speech with precise timing. If humans do rely on some form of coupled oscillation model, as proposed by Wilson and Wilson (2005), this could possibly be used to put the system in concert with the user.

## Declaration of interests

Gabriel Skantze is Professor in speech technology at the Department of Speech Music and Hearing at KTH. He is also co-founder of the company Furhat Robotics.

## Acknowledgments

This work is supported by the Swedish research council (VR) project Coordination of Attention and Turn-taking in Situated Interaction (#2013-1403). Thanks to Nigel Ward for initial discussions on this review, and to Erik Ekstedt, Martin Johansson and Raveesh Meena for their contributions to our work on modelling turn-taking at KTH. Thanks also to the reviewers and editor for their very helpful remarks.

## References

Al Moubayed, S., Edlund, J., Beskow, J., 2012. Taming mona lisa : communicating gaze faithfully in 2D and 3D facial projections. ACM Trans. Interact. Intell. Syst. 1 (2), 1–25. https://doi.org/10.1145/2070719.2070724.

Al Moubayed, S., Skantze, G., 2011. Turn-taking control using gaze in multiparty human-Computer dialogue: effects of 2D and 3D displays. Proceedings of the International Conference on Audio-Visual Speech Processing, 99−102.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R., 1991. The HCRC map task corpus. Lang Speech 34 (4), 351–366.

Andrist, S., Leite, I., Lehman, J., 2013. Fun and fair: influencing turn-taking in a multi-party game with a virtual agent. Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13, 352−355. 10.1145/2485760.2485800

Argyle, M., Cook, M., 1976. Gaze and Mutual Gaze. Cambridge: Cambridge University Press.

Argyle, M., Graham, J.A., 1976. The central europe experiment: looking at persons and looking at objects. Environ. Psychol. Nonverbal Behav. 1 (1), 6–16.

Auer, P., 2018. Gaze, Addressee Selection and Turn-taking in Three-party Interaction. In: Brône, G., Oben, B. (Eds.), Eye-tracking in Interaction: Studies on the Role of Eye Gaze in Dialogue. John Benjamins, pp. 197–232. https://doi.org/10.1075/ais.10.09aue.

Ba, S.O., Odobez, J.-M., 2009. Recognizing visual focus of attention from head pose in natural meetings. IEEE Trans. Syst. Man Cybern. Part B: Cybern. 39 (1), 16–33.

Ball, P., 1975. Listeners' Responses to filled pauses in relation to floor apportionment. Br. J. Soc. Clin. Psychol. 14 (4), 423–424.

Bavelas, J.B., Coates, L., Johnson, T., 2002. Listener responses as a collaborative process: the role of eye gaze. J. Commun. 52 (September), 566–580.

Bell, L., Boye, J., Gustafson, J., 2001. Real-time Handling of Fragmented Utterances. In: Proceedings of the NAACL Workshop on Adaption in Dialogue Systems.

Bennett, A., 1981. Interruptions and the interpretation of conversation. Discourse Process 4, 171–188.

Bögels, S., Torreira, F., 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. J. Phon. 52, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004.

Bohus, D., Horvitz, E., 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In: Proceedings of International Conference on Multimodal Interfaces, ICMI.Beijing, China

Bohus, D., Horvitz, E., 2011. Decisions about turns in multiparty conversation: from perception to action. In: Proceedings of International Conference on Multimodal Interfaces, ICMI, pp. 153–160.

Brady, P.T., 1965. A technique for investigating on off patterns of speech. Bell Syst. Tech. J. 44 (1), 1–22. https://doi.org/10.1002/j.1538-7305.1965.tb04135.x.

Brady, P.T., 1968. A statistical analysis of on off patterns in 16 conversations. Bell Syst. Tech. J. 47 (1), 73–91. https://doi.org/10.1002/j.1538-7305.1968.tb00031.x.

Calisgan, E., Haddadi, A., Loos, H.F.M.V.D., Alcazar, J.A., Croft, E.A., 2012. Identifying nonverbal cues for automated human-robot turn-taking. In: Proceedings of the IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., Yan, H., 2001. Human Conversation as a System Framework: Designing Embodied Conversational Agents. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), Embodied Conversational Agents. MIT Press, Cambridge, MA, US, pp. 29–63.

Clancy, P.M., Thompson, S.A., Suzuki, R., Tao, H., 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. J. Pragmat. 26 (3), 355–387. https://doi.org/10.1016/0378-2166(95)00036-4.

Clark, H., 1996. Using Language. Cambridge University Press, Cambridge, UK.

Clark, H., 2005. Coordinating with each other in a material world. Discourse Stud. 7 (4−5), 507–525.

Clark, H., Fox Tree, J.E., 2002. Using uh and UM in spontaneous speaking. Cognition 84 (1), 73–111.

Coates, J., 1994. No gap, lots of overlap; turn-taking patterns in the talk of women friends. Researching Language and Literacy in Social Context: A Reader. Multilingual Matters.

Cummins, F., 2012. Gaze and blinking in dyadic conversation: a study in coordinated behaviour among individuals. Lang Cogn. Process.. https://doi.org/10.1080/01690965.2011.615220.

De Kok, I., Heylen, D., 2009. Multimodal end-of-turn prediction in multi-party meetings. In: Proceedings of International Conference on Multimodal Interfaces, ICMI, pp. 91–97. https://doi.org/10.1145/1647314.1647332.

Dethlefs, N., Hastie, H., Cuayahuitl, H., Yu, Y., Rieser, V., Lemon, O., 2016. Information density and overlap in spoken dialogue. Comput. Speech Lang. 37, 82–97. https://doi.org/10.1016/j.csl.2015.11.001.

DeVault, D., Sagae, K., Traum, D., 2009. Can I Finish? Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 11–20.London, UK

Dittman, A.T., Llewellyn, L.G., 1967. The phonemic clause as a unit of speech decoding. J. Pers. Soc. Psychol. 6 (3), 341–349. https://doi.org/10.1037/h0024739.

Duncan, S., 1972. Some signals and rules for taking speaking turns in conversations. J. Pers. Soc. Psychol. 23 (2), 283–292.

Duncan, S., 1974. On signalling that it's your turn to speak. J. Exp. Soc. Psychol. 10 (3), 234–247.

Duncan, S., Fiske, D., 1977. Face-to-face Interaction: Research, Methods and Theory. Lawrence Erlbaum Associates, Hillsdale, New Jersey, US.

Edlund, J., Beskow, J., 2009. Mushypeek : A Framework for online investigation of audiovisual dialogue phenomena. Lang Speech 52 (2/3), 351–367. https://doi.org/10.1177/0023830909103179.

Edlund, J., Heldner, M., 2005. Exploring prosody in interaction control. Phonetica 62 (2−4), 215–226. https://doi.org/10.1159/000090099.

Ekstedt, E., Skantze, G., 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2981–2990. https://doi.org/10.18653/v1/2020.findings-emnlp.268.

Ervin-Tripp, S.M., 1979. Children's Verbal Turn-taking. In: Ochs, E., Schieffelin, B. (Eds.), Developmental pragmatics. Academic Press, pp. 391–414.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the ACM Multimedia, pp. 1459–1462.Florence, Italy

Ferrer, L., Shriberg, E., Stolcke, A., 2002. Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody. In: Procedings of the International Conference on Spoken Language Processing, ICSLP, pp. 2061–2064.

Ford, C., Thompson, S., 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: Ochs, E., Schegloff, E., Thompson, A. (Eds.), Interaction and Grammar. Cambridge University Press, Cambridge, pp. 134–184.

French, P., Local, J., 1983. Turn-competitive incomings. J. Pragmat.. https://doi.org/10.1016/0378-2166(83)90147-9.

Gao, Y., Mishchenko, Y., Shah, A., Matsoukas, S., Vitaladevuni, S., 2020. Towards data-efficient modeling for wake word spotting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE, pp. 7479–7483.

Garrod, S., Pickering, M.J., 2015. The use of content and timing to predict turn transitions. Front Psychol. 6, 751. https://doi.org/10.3389/fpsyg.2015.00751.

Gervits, F., Thielstrom, R., Roque, A., Scheutz, M., 2020. It's About Time : Turn-Entry Timing For Situated Human-Robot Dialogue. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 86–96.

Godfrey, J., Holliman, E., McDaniel, J., 1995. Switchboard,: Telephone Speech Corpus for Research and Development. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 517–520.San Francisco

Goffman, E., 1979. Footing. Semiotica 25 (1–2), 1–30.

Goodwin, C., 1981. Conversational Organization: Interaction Between Speakers and Hearers. Academic Press, New York.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. 25 (3), 601–634.

Gravano, A., Hirschberg, J., 2012. A Corpus-Based Study of Interruptions in Spoken Dialogue. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Heins, R., Franzke, M., Durian, M., Bayya, A., 1997. Turn-taking as a design principle for barge-in spoken language systems. Int. J. Speech Technol. 2 (2), 155–164.

Heldner, M., Edlund, J., 2010. Pauses, gaps and overlaps in conversations. J Phon 38 (4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002.

Hemphill, C.T., Godfrey, J.J., Doddington, G.R., 1990. The atis spoken language systems pilot corpus. In: Proceedings of Speech and Natural Language Workshop.

Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice. Speech Commun. 53 (1), 23–35. https://doi.org/10.1016/j.specom.2010.08.003.

Hjalmarsson, A., Oertel, C., 2011. Gaze direction as a backchannel inviting cue in dialogue. In: Proceedings of the IVA 2012 workshop on Realtime Conversational Virtual Agents, p. 1.Santa Cruz, CA

Holler, J., Kendrick, K.H., Levinson, S.C., 2018. Processing language in face-to-face conversation: questions with gestures get faster responses. Psychonomic Bull. Rev. 25 (5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z.

Hough, J., Schlangen, D., 2016. Investigating Fluidity for Human-Robot Interaction with Real-time, Real-world Grounding Strategies. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 288–298. https://doi.org/10.18653/v1/W16-3637.

Hussain, N., Erzin, E., Metin Sezgin, T., Yemez, Y., 2019. Speech driven backchannel generation using deep Q-network for enhancing engagement in human-robot interaction. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. https://doi.org/10.21437/Interspeech.2019-2521.

Ishii, R., Otsuka, K., Kumano, S., Yamato, J., 2014. Analysis of respiration for prediction of "Who Will Be Next Speaker and When?" in multi-party meetings. In: Proceedings of the International Conference on Multimodal Interfaces, ICMI. https://doi.org/10.1145/2663204.2663271.

Ishii, R., Otsuka, K., Kumano, S., Yamato, J., 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. ACM Trans. Interact. Intell. Syst.. https://doi.org/10.1145/2757284.

Johansson, M., Hori, T., Skantze, G., Hïthker, A., Gustafson, J., Höthker, A., Gustafson, J., 2016. Making turn-taking decisions for an active listening robot for memory training. In: Proceedings of the International Conference on Social Robotics, 9979 LNAI, pp. 940–949. https://doi.org/10.1007/978-3-319-47437-3_92.

Johansson, M., Skantze, G., 2015. Opportunities and Obligations to take turns in collaborative multi-party human-robot interaction. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, p. 305314.

Johansson, M., Skantze, G., Gustafson, J., 2013. Head pose patterns in multiparty human-robot team-building interactions. In: Proceedings of International Conference on Social robotics (ICSR), 8239 LNAI, . https://doi.org/10.1007/978-3-319-02675-6_35.

Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S., 2013. Gaze and turn-taking behavior in casual conversational interactions. ACM Trans. Interact. Intell. Syst. 3 (2).

Jokinen, K., Nishida, M., Yamamoto, S., 2010. On Eye-gaze and Turn-taking. In: Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 118–123. https://doi.org/10.1145/2002333.2002352.

Jonsdottir, G.R., Thorisson, K.R., Nivel, E., 2008. Learning smooth, human-like turntaking in realtime dialogue. In: Proceedings of the Intelligent Virtual Agents, IVA. https://doi.org/10.1007/978-3-540-85483-8_17.

Jovanovic, N., Op Den Akker, R., Nijholt, A., 2006. Addressee identification in face-to-face meetings. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL, pp. 169–176.

Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., Waibel, A., 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In: Proceedings of International Conference on Multimodal Interfaces, ICMI.PA, USA

Kawahara, T., Iwatate, T., Takanashi, K., 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations.. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Kendon, A., 1967. Some functions of gaze direction in social interaction. Acta Psychol. (Amst) 26, 22–63.

Khouzaimi, H., Laroche, R., Lefèvre, F., 2015. Optimising turn-taking strategies with reinforcement learning. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 315–324. https://doi.org/10.18653/v1/w15-4643.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y., 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Lang Speech. 41, 295–321. https://doi.org/10.1177/002383099804100404.

Kunc, L., Míkovec, Z., Slavík, P., 2013. Avatar and dialog turn-yielding phenomena. Int. J. Technol. Hum. Interact. 9 (2), 66–88. https://doi.org/10.4018/jthi.2013040105.

Lala, D., Inoue, K., Kawahara, T., 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In: Proceedings of International Conference on Multimodal Interfaces, ICMI, pp. 226–234. https://doi.org/10.1145/3340555.3353727.

Lee, C., Narayanan, S., 2010. Predicting interruptions in dyadic spoken interactions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5250–5253.

Levelt, W., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, Mass., USA.

Levinson, S.C., Torreira, F., 2015. Timing in turn-taking and its implications for processing models of language. Front. Psychol. 6 (JUN). https://doi.org/10.3389/fpsyg.2015.00731.

Leyzberg, D., Spaulding, S., Toneva, M., Scassellati, B., 2012. The physical presence of a robot tutor increases cognitive learning gains. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society.ISBN 978-0-9768318-8-4

Local, J., Kelly, J., Wells, W., 1986. Towards a phonology of conversation: turn-taking in tyneside english. J. Linguist. 22 (2), 411–437.

Maier, A., Hough, J., Schlangen, D., 2017. Towards deep end-of-Turn prediction for situated spoken dialogue systems. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017, pp. 1676–1680. https://doi.org/10.21437/Interspeech.2017-1593.

Martinez, M.A., 1987. Dialogues among children and between children and their mothers. Child Dev. 58 (4), 1035–1043. https://doi.org/10.2307/1130544.

Masumura, R., Tanaka, T., Ando, A., Ishii, R., Higashinaka, R., Aono, Y., 2018. Neural dialogue context online end-of-turn detection. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pp. 224–228. https://doi.org/10.18653/v1/W18-5024.Melbourne, Australia

McFarland, D.H., 2001. Respiratory markers of conversational interaction. J. Speech Lang. Hear. Res. 44, 128–143.

McGlashan, S., Burnett, D. C., Carter, J., Danielsen, P., Ferrans, J., Hunt, A., Lucas, B., Porter, B., Rehor, K., Tryphonas, S., 2004. Voice extensible markup language (VoiceXML): version 2.0.

Meena, R., Skantze, G., Gustafson, J., 2014. Data-driven models for timing feedback responses in a map task dialogue system. Comput. Speech Lang. 28 (4), 903–922.

Mondada, L., 2007. Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. Discourse Stud.. https://doi.org/10.1177/1461445607075346.

Moon, A., Troniak, D., Gleeson, B., Pan, M., Zheng, M., Blumer, B., MacLean, K., Croft, E., 2014. Meet Me Where I'm Gazing: How Shared Attention Gaze Affects Human-robot Handover Timing. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, HRI, pp. 334–341. https://doi.org/10.1145/2559636.2559656.New York, NY

Morency, L.P., de Kok, I., Gratch, J., 2008. Predicting listener backchannels: A probabilistic multimodal approach. In: Proceedings of the Intelligent Virtual Agents, IVA. Springer, Tokyo, Japan, pp. 176–190.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H., 2012. Conversational gaze mechanisms for humanlike robots. ACM Trans. Interact. Intell. Syst. 1 (2), 1–12. https://doi.org/10.1145/2070719.2070725.

Nakano, Y.I., Yoshino, T., Yatsushiro, M., Takase, Y., 2015. Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. ACM Trans. Interact. Intell. Syst. 5 (4), 1–23. https://doi.org/10.1145/2743028.

Neiberg, D., Truong, K.P., 2011. Online detection of vocal Listener Responses with maximum latency constraints. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5836–5839. https://doi.org/10.1109/ICASSP.2011.5947688.

O'Conaill, B., Whittaker, S., Wilbur, S., 1993. Conversations over video conferences: an evaluation of the spoken aspects of video-Mediated communication. Human-Comput. Int.. https://doi.org/10.1207/s15327051hci0804_4.

O'Connell, D.C., Kowal, S., 2005. Uh and um revisited: are they interjections for signaling delay? J. Psychol. Res. 34 (6), 555–576. https://doi.org/10.1007/s10936-005-9164-3.

O'Connell, D.C., Kowal, S., Kaltenbacher, E., 1990. Turn-taking: a critical analysis of the research tradition. J. Psychol. Res. 19 (6), 345–373. https://doi.org/10.1007/BF01068884.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., Gustafson, J., 2012. Gaze patterns in turn-taking. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Oertel, C., Wlodarczak, M., Tarasov, A., Campbell, N., Wagner, P., 2012. Context cues for classification of competitive and collaborative overlaps. In: Proceedings of the International Conference on Speech Prosody, pp. 721–724.

Pelachaud, C., Badler, N., Steedman, M., 1996. Generating facial expressions for speech. Cogn. Sci. 20 (1).

Poesio, M., Rieser, H., 2010. Completions, Coordination, and Alignment in Dialogue. Dial. Discour. 1 (1). https://doi.org/10.5087/dad.2010.001.

Raux, A., Bohus, D., Langner, B., Black, A.W., Eskenazi, M., 2006. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.Pittsburgh, PA, USA

Raux, A., Eskenazi, M., 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL.Columbus, OH, USA

Raux, A., Eskenazi, M., 2009. A finite-state turn-taking model for spoken dialog systems. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, pp. 629–637.Boulder, CO, USA

Rochet-Capellan, A., Fuchs, S., 2014. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. Philosoph. Trans. R. Soc. B Biol. Sci.. https://doi.org/10.1098/rstb.2013.0399.

Roddy, M., Harte, N., 2020. Neural Generation of Dialogue Response Timings. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL.arXiv:2005.09128v1

Roddy, M., Skantze, G., Harte, N., 2018. Investigating speech features for continuous turn-taking prediction using LSTMs. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.Hyderabad, India

Roddy, M., Skantze, G., Harte, N., 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. In: Proceedings of the International Conference on Multimodal Interfaces, ICMI, pp. 186–190. https://doi.org/10.1145/3242969.3242997.New York, New York, USA

Ruede, R., Müller, M., Stüker, S., Waibel, A., 2019. Yeah, right, uh-huh: A deep learning backchannel predictor. In: Proceedings of the Lecture Notes in Electrical Engineering. https://doi.org/10.1007/978-3-319-92108-2_25.

de Ruiter, J.-P., Mitterer, H., Enfield, N.J., 2006. Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. Language (Baltim) 82 (3), 515–535. https://doi.org/10.1353/lan.2006.0130.

Sacks, H., Schegloff, E., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language (Baltim) 50, 696–735.

Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., Aikawa, K., 2002. Learning decision trees to determine turn-taking by spoken dialogue systems. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP.

Schegloff, E., 2000. Overlapping talk and the organization of turn-taking for conversation. Lang. Soc. 29 (1), 1–63.

Schlangen, D., 2006. From reaction to prediction: experiments with computational models of turn-taking. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2010–2013.Pittsburgh, PA, USA

Schlangen, D., Skantze, G., 2009. A general, abstract model of incremental dialogue processing. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL, pp. 710–718.

Schneider, L., Goffman, E., 1964. Behavior in public places: notes on the social organization of gatherings.. Am. Sociol Rev.. https://doi.org/10.2307/2091496.

Selfridge, E.O., Arizmendi, I., Heeman, P.A., Williams, J.D., 2013. Continuously predicting and processing barge-in during a live spoken dialogue task. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 384–393.

Selfridge, E.O., Heeman, P.a., 2010. Importance-driven turn-bidding for spoken dialogue systems. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, pp. 177–185.Uppsala, Sweden

Sellen, A.J., 1995. Remote conversations: the effects of mediating talk with technology. Human-Comput. Int.. https://doi.org/10.1207/s15327051hci1004_2.

Selting, M., 1996. On the interplay of syntax and prosody in the constitution of turnconstructional units and turns in conversation. Pragmatics 6, 357–388.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Heck, L., 2012. Learning when to listen: detecting system-addressed speech in human-human-computer dialog. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.Partland, OR, USA

Sikveland, R.O., Ogden, R., 2012. Holding gestures across turns. Gesture 12 (2), 166–199. https://doi.org/10.1075/gest.12.2.03sik.

Skantze, G., 2016. Real-Time coordination in human-Robot interaction using face and voice. AI Magazine 37 (4), 19. https://doi.org/10.1609/aimag.v37i4.2686.

Skantze, G., 2017. Predicting and regulating participation equality in human-robot conversations: effects of age and gender. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, HRI, pp. 196–204. https://doi.org/10.1145/2909824.3020210.

Skantze, G., 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 220–230. https://doi.org/10.18653/v1/W17-5527.

Skantze, G., Gustafson, J., 2009. Attention and interaction control in a human-human-computer dialogue setting. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL.

Skantze, G., Hjalmarsson, A., 2013. Towards incremental speech generation in conversational systems. Comput. Speech Lang. 27 (1), 243–262.

Skantze, G., Hjalmarsson, A., Oertel, C., 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. Speech Commun. 65, 50–66. https://doi.org/10.1016/j.specom.2014.05.005.

Skantze, G., Johansson, M., Beskow, J., 2015. Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. In: Proceedings of the ACM on International Conference on Multimodal Interaction, pp. 67–74. https://doi.org/10.1145/2818346.2820749.

Skantze, G., Schlangen, D., 2009. Incremental dialogue processing in a micro-domain. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL, pp. 745–753.

Stiefelhagen, R., Yang, J., Waibel, A., 2002. Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans. Neural Networks 13 (4), 928–938.

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.-P., Yoon, K.-E., Levinson, S.C., 2009. Universals and cultural variation in turn-taking in conversation.. Proc. Natl. Acad. Sci. U.S.A. 106 (26), 10587–10592. https://doi.org/10.1073/pnas.0903616106.

Streeck, J., Hartge, U., 1992. Previews: gestures at the transition place. In: Auer, P., di Luzio, P.A. (Eds.), The Contextualization of Language. Amsterdam: Benjamins, pp. 135–157.

Strohkorb, S., Leite, I., Warren, N., Scassellati, B., 2015. Classification of children's social dominance in group interactions with robots. In: Proceedings of the International Conference on Multimodal Interfaces, ICMI, pp. 227–234.

Ström, N., Seneff, S., 2000. Intelligent barge-in conversational systems. In: Proceedings of the International Conference on Spoken Language Processing, ICSLP.

Strömbergsson, S., Hjalmarsson, A., Edlund, J., House, D., 2013. Timing responses to questions in dialogue. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2584–2588.

Ten Bosch, L., Oostdijk, N., Boves, L., 2005. On temporal aspects of turn taking in conversational dialogues. Speech Commun. 47. https://doi.org/10.1016/j.specom.2005.05.009.

Thórisson, K.R., 1999. A mind model for multimodal communicative creatures and humanoids. Int. J. Appl. Artif. Intell. 13 (4–5), 449–486.

Tomasello, M., Hare, B., Lehmann, H., Call, J., 2007. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. J. Hum. Evol. 52 (3), 314–320.

Torreira, F., Bögels, S., Levinson, S.C., 2015. Breathing for answering: the time course of response planning in conversation. Front Psychol. 6. https://doi.org/10.3389/fpsyg.2015.00284.

Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., Swartout, W., 2012. Ada and grace: Direct interaction with museum visitors. In: Proceedings of the International Conference on Intelligent Virtual Agents. Springer, pp. 245–251.

Traum, D., Rickel, J., 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In: Proceedings of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2, pp. 766–773. https://doi.org/10.1145/544862.544922.New York, NY, USA

Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., Vaswani, A., 2007. Hassan: A virtual human for tactical questioning. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 71–74.

Truong, K., Poppe, R., Heylen, D., 2010. A rule-based backchannel prediction model using pitch and pause information.. In: Kobayashi, T., Hirose, K., Nakamura, S. (Eds.), Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3058–3061.

Truong, K.P., 2013. Classification of cooperative and competitive overlaps in speech using cues from the context,overlapper, and overlappee. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1404–1408.

Velichkovsky, B.M., 1995. Communicating attention: gaze position transfer in cooperative problem solving. Pragmat. Cogn. 3, 199224.

Vinyals, O., Bohus, D., Caruana, R., 2012. Learning speaker, addressee and overlap detection models from multimodal streams. In: Proceedings of the 14th ACM international conference on Multimodal Interaction. ACM, pp. 417–424.

Wang, Y.T., Nip, I.S., Green, J.R., Kent, R.D., Kent, J.F., Ullman, C., 2012. Accuracy of perceptual and acoustic methods for the detection of inspiratory loci in spontaneous speech. Behav. Res. Methods 44 (4), 1121–1128. https://doi.org/10.3758/s13428-012-0194-0.

Ward, N., 1996. Using prosodic clues to decide when to produce backchannel utterances. In: Proceedings of the fourth International Conference on Spoken Language Processing, pp. 1728–1731.Philadelphia, USA

Ward, N., 2004. Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In: Proceedings of the International Conference on Speech Prosody, pp. 325–328.10.1.1.2.433

Ward, N., 2019. Prosodic Patterns in English Conversation. Cambridge University Press. https://doi.org/10.1017/9781316848265.

Ward, N., Aguirre, D., Cervantes, G., Fuentes, O., 2018. Turn-Taking Predictions across Languages and Genres Using an LSTM Recurrent Neural Network. In: Proceedings of the IEEE Spoken Language Technology Workshop, SLT, pp. 831–837. https://doi.org/10.1109/SLT.2018.8639673.

Ward, N., Rivera, A., Ward, K., Novick, D., 2005. Root causes of lost time and user stress in a simple dialog system. In: Proceedings of Interspeech 2005.Lisbon, Portugal

Weiss, C., 2018. When gaze-selected next speakers do not take the turn. J. Pragmat.. https://doi.org/10.1016/j.pragma.2018.05.016.

Weizenbaum, J., 1966. ELIZA - A computer program for the study of natural language communication between man and machine. Commun. Assoc. Comput. Mach. 9, 36–45.

Wilson, M., Wilson, T.P., 2005. An oscillator model of the timing of turn-taking. Psychon. Bull. Rev. 12 (6), 957–968.

Witt, S., 2015. Modeling user response timings in spoken dialog systems. Int. J. Speech Technol. 18 (2), 231–243. https://doi.org/10.1007/s10772-014-9265-1.

Włodarczak, M., Heldner, M., 2016. Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 510–514. https://doi.org/10.21437/Interspeech.2016-344.

Woodruff, A., Aoki, P.M., 2003. How push-to-talk makes talk less pushy. In: Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work, pp. 170–179. https://doi.org/10.1145/958160.958187.New York, NY, USA

Yang, L.-C., 2001. Visualizing spoken discourse: prosodic form and discourse functions of interruptions. In: Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL, pp. 1–10. https://doi.org/10.3115/1118078.1118106.

Yngve, V.H., 1970. On getting a word in edgewise. In: Proceedings of the Papers from the sixth regional meeting of the Chicago Linguistic Society. Department of Linguistics, Chicago, pp. 567–578.

Zellers, M., House, D., Alexanderson, S., 2016. Prosody and hand gesture at turn boundaries in Swedish. In: Proceedings of the International Conference on Speech Prosody, pp. 831–835. https://doi.org/10.21437/speechprosody.2016-170.

Zima, E., Weiß, C., Brône, G., 2019. Gaze and overlap resolution in triadic interactions. J. Pragmat. 140, 49–69. https://doi.org/10.1016/j.pragma.2018.11.019.