frontiers in Computer Science

REVIEW
published: 15 March 2022
doi: 10.3389/fcomp.2022.744574

Modeling Feedback in Interaction With Conversational Agents—A Review

Agnes Axelsson[1†], *Hendrik Buschmeier*[2*†] *and Gabriel Skantze*[1†]

[1] Division of Speech, Music and Hearing (TMH), KTH Royal Institute of Technology, Stockholm, Sweden, [2] Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Intelligent agents interacting with humans through conversation (such as a robot, embodied conversational agent, or chatbot) need to receive feedback from the human to make sure that its communicative acts have the intended consequences. At the same time, the human interacting with the agent will also seek feedback, in order to ensure that her communicative acts have the intended consequences. In this review article, we give an overview of past and current research on how intelligent agents should be able to both give meaningful feedback toward humans, as well as understanding feedback given by the users. The review covers feedback across different modalities (e.g., speech, head gestures, gaze, and facial expression), different forms of feedback (e.g., backchannels, clarification requests), and models for allowing the agent to assess the user's level of understanding and adapt its behavior accordingly. Finally, we analyse some shortcomings of current approaches to modeling feedback, and identify important directions for future research.

Keywords: feedback, grounding, spoken dialogue, multimodal signals, human-agent interaction, review

OPEN ACCESS

Edited by:
Kostas Karpouzis,
Panteion University, Greece

Reviewed by:
Christos Troussas,
University of West Attica, Greece
Michael McTear,
Ulster University, United Kingdom
Costanza Navarretta,
University of Copenhagen, Denmark

***Correspondence:**
Hendrik Buschmeier
hbuschme@uni-bielefeld.de

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *20 July 2021*
Accepted: *11 February 2022*
Published: *15 March 2022*

Citation:
Axelsson A, Buschmeier H and Skantze G (2022) Modeling Feedback in Interaction With Conversational Agents—A Review. Front. Comput. Sci. 4:744574. doi: 10.3389/fcomp.2022.744574

1. INTRODUCTION

Any intelligent system interacting with an environment needs to receive **feedback** in order to understand the consequences of its actions (Wiener, 1948). Thus, an intelligent system interacting with humans (such as a robot, virtual agent, or chatbot) also needs to receive feedback across multiple modalities to make sure that its communicative acts have the intended consequences. At the same time, the human interacting with the agent will also seek feedback, in order to ensure that her communicative acts have the intended consequences. Another way of saying this is that the agent and the human need to reach **mutual understanding**, which means that both parties believe that they share common ground (Clark, 1996), and in this process, the exchange of feedback is vital (Allwood et al., 1992).

Clark (1996) proposes that communication depends on a two-track system. On **Track 1**, the main communicative goals are accomplished (such as getting an answer to a question or proposing a joint dinner). On **Track 2**, the speakers exchange feedback regarding their communicative success. Thus, every contribution enacts the collateral question "Do you understand what I mean by this?". This model is illustrated in **Figure 1**, which also illustrates the different tasks that have to be considered when developing an agent that is both capable of **generating** feedback toward the human, as well as **identifying** and **classifying** feedback received from the human.

In this review article, we will provide an overview of how feedback in human-agent conversation has been modelled. There already exist a few review articles that focus on how agents should be able

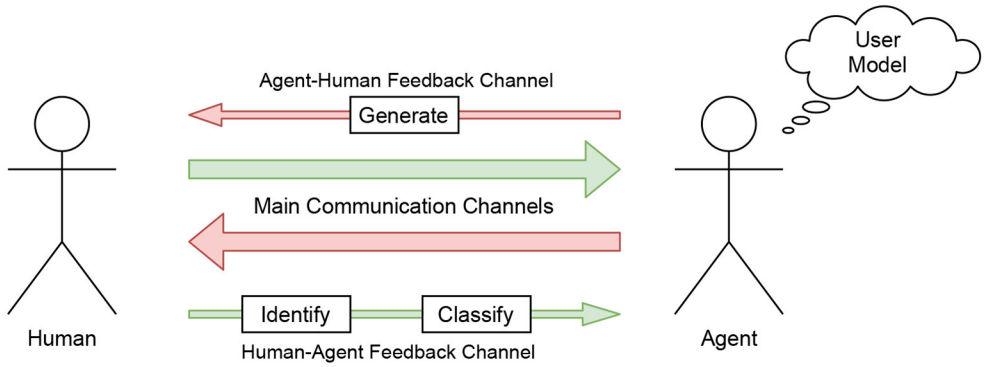

FIGURE 1 | An abstract representation of how feedback-aware and feedback-generating systems function.

to produce feedback in the form of *backchannels* (brief listener responses; de Kok and Heylen, 2012; Bevacqua, 2013). However, we have noticed a lack of reviews providing a more holistic overview of the area, which do not just focus on backchannels, and which take both human-agent and agent-human feedback into account. This is the motivation for this review.

Although feedback is relevant for any form of human-machine interaction, we will focus on conversational interaction in this review. We will cover research related to the interaction between humans, robots, virtual agents, and chatbots, but we will put less stress on the specific platform or embodiment of the agent, as we believe most of the aspects related to feedback should be of generic interest. Another theme in this review is to highlight the theories on feedback that come from studies of human-human conversation. We hope this can motivate a higher awareness of such theories in the development of computational models of feedback in human-agent conversation.

This review article is structured as follows. We will start with a general background on theories of feedback in conversation between humans (Section 2). We will then focus on research done on modeling feedback produced by the agent toward the human (Section 3), followed by coverage of work done on feedback from the human toward the agent (Section 4). After this, we will provide an analysis of the research field in order to highlight topics where we think more research is needed (Section 5).

2. FEEDBACK IN COMMUNICATION

The term feedback comes originally from the field of cybernetics, defined by Wiener (1948) as "the scientific study of control and communication in the animal and the machine." There, it denotes the general processes by which a control unit gets information about the consequences of its actions. Thus, in linguistics, feedback should be understood as the more specific process by which speakers get information about how their communicative act was received by their listeners, and the consequences it may have.

A general distinction can be made between *negative* and *positive* feedback (Allwood et al., 1992; Clark, 1996), where negative feedback informs the speaker that the communicative act did not have the expected consequences, and positive feedback informs the speaker that it did. The earlier the speaker receives feedback, the sooner any problems in the communication can be addressed. Thus, speakers do not typically wait to receive feedback until after their contribution is complete. Instead, they continuously monitor the addressee for understanding and may alter and adapt their utterance as it is being produced (Bavelas et al., 2000; Clark and Krych, 2004).

One way of investigating the importance of feedback in spoken interaction is to compare an interactive setting, where one person gives instructions to another, to a pre-recorded (non-interactive) instruction. Such studies have shown that non-interactive settings (which lack opportunity for feedback) result in the production of longer and less intelligible referring expressions (Krauss and Weinheimer, 1966; Clark and Krych, 2004). Bavelas et al. (2000) investigated the setting of one person telling a story to another person. In an experimental condition where listeners were distracted, they produced fewer feedback responses, which in turn made the narrator tell the story less well. In this sense, the listeners could be described as "co-narrators," and the results highlight the importance of moment-by-moment feedback in conversation.

According to Clark (1996), communication can be described as the process by which we make our knowledge and beliefs common, we add to our **common ground**, which should be understood as the sum of our mutual knowledge, beliefs, and suppositions. The process by which this is accomplished is called *grounding*, which involves both a *presentation phase* and an *acceptance phase*, corresponding to the two tracks illustrated in **Figure 1**. In these terms, the feedback is found in the acceptance phase. However, it is important to stress that most contributions involve both an acceptance/feedback (of what was just said) as well as a presentation of something new (Clark and Schaefer, 1989).

2.1. Feedback on Different Levels of Action

Both Allwood et al. (1992) and Clark (1996) make a distinction between four levels of action that take place when a speaker is

TABLE 1 | Examples of positive and negative feedback on different levels and across different modalities (e.g., speech, facial expression, head gesture).

Level	Positive	Negative
Contact	Backchannel ("mhm," nod)	"Are you there?"
Perception	Backchannel,	Repair initiator ("huh?"), Frown
Understanding	Reprise fragment ("blue")	Clarification request ("blue?"), Frown
Attitude/Acceptance	Acknowledgement ("okay"), Agreement ("I agree"), Smile	"I don't agree," "I cannot find that"

trying to communicate something to a listener, and feedback can be related to these different levels. According to Allwood et al. (1992), feedback can be related to:

- *Attitude*: The listener's attitude toward the message. This could involve whether they accept or reject a statement as being true, or are willing to answer a question or accept a proposal, but also emotional attitudes (e.g., whether they like or dislike the message, or find it fun or boring).
- *Understanding*: Whether the listener is able to understand the message.
- *Perception*: Whether the listener is able to perceive the message.
- *Contact*: Whether the listener is willing and able to continue the interaction (e.g., whether they pay attention to the speaker).

According to Clark (1996), for communication to be "successful" between two interlocutors, all these levels of action must succeed. The order of the levels is important: in order to succeed on one level, all the levels below it must be completed. Thus, we cannot understand what a person is saying without hearing the words spoken, we cannot hear the words without attending, and so on. Clark (1996) calls this the principle of *upward completion*. As communication problems may arise on all these levels, positive and negative feedback can be given on a specific level. For example, the phrase "sorry, what did you say" gives negative feedback specifically on the level of perception. By the principle of *downward evidence*, when positive evidence is given on one level, all the levels below it are considered complete. Therefore, the phrase "Okay, I see" entails not just positive acceptance, but also positive understanding, perception and contact. Some examples of positive and negative feedback on the four different levels are shown in **Table 1**.

Of course, we do not always provide positive feedback on every piece of information received. To some extent we must assume understanding and acceptance (as long as we do not get negative feedback), or else communication would not be very efficient. Whether we require positive feedback is dependent on the current situation and task. Clark (1996) uses the term *grounding criterion* to denote this. If the cost of misunderstanding is very high and has irreversible effects (for example if A asks B to delete a file on their computer), the grounding criterion is high, and both A and B are likely to exchange a lot of feedback to ensure common ground before the command is executed. If it is not so high (for example if A asks B to pass the milk), the feedback can be omitted for sake of efficiency, leaving potential misunderstandings to be sorted out later on.

2.2. Display of Understanding

Given the grounding criterion, positive feedback can be characterized as stronger or weaker (apart from the four levels outlined above), where stronger feedback is typically less efficient, as discussed by Clark (1996). A simple "okay" might indicate that the listener *thinks* that she has understood, without the speaker being able to confirm this. Another form of weak evidence is to simply provide a "relevant next contribution," as in the following example (where the speaker can at least partly confirm that the overall intent was understood):

A: I want to go to Paris
B: On which date would you like to go?

A stronger form of positive feedback discussed by Clark (1996) is *display of understanding*, where the listener repeats (or rephrases) parts of the last contribution in their own contribution. While being less efficient, it allows for the speaker to better verify the reception. This can often be mixed with a new initiative, as in the following example (sometimes called *implicit verification* in the dialog system literature):

A: I want to go to Paris
B: On which date would you like to go to Paris?

Using the distinction made by Clark (1996) between *Track 1* and *Track 2* (as mentioned in the Section 1), this illustrates how a contribution can mix information on both these tracks. However, a display of understanding can of course also be done as a contribution of its own:

A: I want to go to Paris
B: Okay, to Paris

2.3. Clarification Requests

A special form of negative feedback are *clarification requests*. These are typically not mixed with the next contribution (i.e., they belong exclusively to Track 2) and thus have to be resolved before the dialogue can proceed. Clarification requests can be formulated as a wh-question ("What did you say?"), a yes/no-question ("Did you say Paris?") or an alternative question ("Did you say Paris or London?"). They can also be given on the different action levels outlined above, such as hearing ("Did you say Paris?") or understanding ("Do you mean Paris, France or Paris, Texas?"; Rodríguez and Schlangen, 2004).

Many clarification requests are also provided in elliptical form, as reprise fragments ("To Paris?"), reprise sluices ("A red what?"), gaps ("A red …?"), or in conventional form ("Huh?" "Pardon?"; Purver, 2004). Thus, a reprise fragment (such as "to Paris?") can be very similar to the display of understanding discussed above, although the interrogative interpretation puts a higher expectation on the speaker to confirm the proposed

interpretation. The difference can be marked prosodically. In English, for example, a rising pitch can indicate that it is a clarification request (Skantze et al., 2006).

2.4. Backchannels

The distinction made by Clark (1996) between Track 1 and Track 2 is similar to, but not exactly the same as, the difference described by Yngve (1970) between the *main channel* and the *backchannel*. So-called **backchannels** can take the form of brief, relatively soft vocalizations (e.g., "mm hm," "uh huh," "yeah") or gestures (e.g., head nods, facial expressions; Yngve, 1970). Whereas, feedback in general can be produced as separate turns (e.g., clarification requests), backchannels do not claim the floor and are thus often produced in overlap with the speaker. The phenomenon has also been referred to as "listener responses" (Dittman and Llewellyn, 1967) and "accompaniment signals" (Kendon, 1967) as well as many other terms (cf. Fujimoto, 2007).

Backchannels are often produced to maintain contact and show continued attention, i.e., positive feedback on the lowest level on the action ladder described above. Thus, the best way to make someone stop speaking (at least over the telephone) is to be completely silent (it will not take long before the other speaker will say "are you there?"). However, backchannels can be ambiguous, since they may also commit to higher action levels, depending on their realization (Shimojima et al., 1998). Small differences in prosody can have a big effect on their perceived meaning, and they can even have a negative function (like a prolonged "yeah …" with a falling pitch). In a perception experiment, Lai (2010) found that different intonation contours of cue words (e.g., "yeah," "right," "really") influence listeners' perception of uncertainty. Gravano et al. (2007) did a similar analysis of the word "okay," and found that both prosody and dialogue context affected the interpretation of the word as either a backchannel, an acknowledgement, or a beginning of a new discourse segment. In a study on human-robot interaction, Skantze et al. (2014) found that both the lexical choice and prosody in the users' feedback are correlated with uncertainty, and built a logistic regression classifier to combine these features. It has also been shown how these different functions can be achieved by varying the prosodic realization when synthesizing short feedback utterances (Wallers et al., 2006; Stocksmeier et al., 2007).

2.5. Timing and Elicitation of Feedback

The timing of feedback is important, as not all points of time in an interaction are equally appropriate for providing feedback (Skantze, 2021). At certain points in time, there are *transition relevance places* (Sacks et al., 1974), where the turn can (but does not have to) shift. Since backchannel feedback is typically not considered to constitute a turn, it does not follow regular turn-taking patterns. However, the timing of such feedback is still coordinated, and should ideally be produced in stretches of time called *backchannel relevance spaces* (Heldner et al., 2013; Howes and Eshghi, 2021). In general, speakers coordinate their turn-taking using turn-yielding and turn-holding cues in different modalities (e.g., falling vs. rising pitch; Skantze, 2021). It is not clear to what extent speakers produce these cues intentionally, but they can certainly be used by the speaker to **elicit feedback** from the listener.

Ward (1996) investigated a corpus of Japanese conversations to find predictive cues for backchannels and found that backchannels tended to come about 200 ms after a region of low pitch. Gravano and Hirschberg (2011) investigated English conversations and found that backchannels were often preceded by rising pitch and higher intensity. Bavelas et al. (2002) also examined gaze behavior around backchannels in dyadic interactions, where one person was telling a story to the other. They found that the speaker gazed 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.

2.6. Feedback in Different Modalities

As exemplified in **Table 1**, feedback can be expressed in different modalities. In addition to verbal and verbal-vocal feedback signals, humans use head and hand gestures, facial expressions, eye gaze, and other bodily means to provide feedback in communication. Embodied listener feedback can be perceived visually, which has the advantage that it "interferes" even less with speakers' ongoing verbal production than verbal-vocal feedback (which is produced in a perceptually unobtrusive way to minimize the potential of it being seen as a turn-taking attempt and thus being disruptive; Heldner et al., 2010). Indeed, non-verbal feedback is more likely to co-occur with speech than verbal-vocal feedback (Truong et al., 2011). Nonverbal feedback often combines two or more nonverbal modalities at once (e.g., a head nod combined with a smile) or nonverbal with verbal-vocal modalities (e.g., a head nod combined with an "uh-huh"). Such multimodal feedback expressions are frequent (Allwood and Cerrato, 2003; Allwood et al., 2007; Malisz et al., 2016).

Gaze can serve as a turn-taking cue (Novick et al., 1996; Jokinen et al., 2013), a backchannel-inviting cue from the speaker toward the listener (Bavelas et al., 2002; Hjalmarsson and Oertel, 2012), and/or an indication of *mutual attention* (Frischen et al., 2007). In terms of feedback, attending to the other speaker corresponds to the lowest level of the feedback ladder discussed in Section 2.1 above. If there is a cooperative task with objects in the shared space, so-called *joint attention* can help to make sure that both speakers attend to the same objects, and that references to those objects are understood (Skantze et al., 2014). Gaze can also be a signal from the listener to inform the speaker that the listener is ready to interpret the speaker's facial expressions (Jokinen and Majaranta, 2013). Nakano et al. (2003) showed that maintaining gaze on the speaker was often a sign of negative grounding, specifically non-understanding.

Heylen (2006, 2008) argues that **head gestures** are an especially important form of feedback when human communication partners can see each other. Heylen (2006) specifically points out three different uses of head gestures as feedback: First, to signal that the listener is processing the proposal of the speaker. An extension of this is to signal that something the speaker said is especially hard to process or take in. Second, to mark that the speaker can take the turn back and

continue speaking, and third, to express an attitudinal reaction to the content presented by the speaker, often in combination with verbal feedback.

Porhet et al. (2017) found that head movements are the most common feedback modality both from doctors toward their patients and vice versa in a medical interaction corpus. The authors also found that a nod from the speaking doctor toward the listening patient is followed by a nod in response from the patient 29% of the time. Włodarczak et al. (2012) found that non-verbal feedback, including head movements, became more common from listeners toward speakers when the listeners were distracted by an unrelated task (pressing a button whenever the speaker spoke a word starting with "s"). Inden et al. (2013) found that head gestures from the listener were increasingly more likely as time passes in the speaker's turn. Forty percent of the listener's head gestures overlapped with the end of the speaker's turn, which indicates that they serve as a turn-yielding cue.

3. FEEDBACK FROM AGENTS TOWARD USERS

This section reviews the state of the art in conversational agents that are able to generate linguistic and multimodal feedback in response to their human interaction partners' utterances. We begin by looking at different motivations for providing backchannel feedback and resulting design and modeling decisions (Section 3.1). We follow up on this by discussing the use of feedback to handle uncertainties stemming from automatic speech recognition (ASR) and contrast this with the research on backchannel-like feedback (Section 3.2). We then review different aspects that need to be considered when providing feedback: timing (Section 3.3), function (Section 3.4), form and multimodality (Section 3.5). Finally, we conclude the section with implications for future conversational agent development (Section 3.6).

3.1. Motivations for Agents to Providing Feedback

As we have seen, feedback in dialogue has several purposes and the motivation for endowing agents with mechanisms to provide feedback to human users varies among authors. A rather practically oriented motivation, mentioned in the literature early on, is based on the importance of feedback as a "design principle" in human-computer interaction in general (Norman, 1990). Processing user input takes its time and a spoken language system that does not provide feedback on (at least) the levels of contact or perception leaves the user in an "ambiguous silence" (Yankelovich et al., 1995) that can either mean that processing of user speech is still ongoing or that user input was not even perceived and waiting longer will not solve the problem. A system that provides feedback makes its state of processing transparent. Users can evaluate immediately whether an utterance was perceived and act on that information (e.g., by waiting or by making another attempt). Feedback of this type has been widely adopted in voice assistants, where it is often displayed visually (using light indicators or on-screen visualizations) or auditorily (using non-linguistic sound effects). Embodied, anthropomorphic agents can use natural human-like signals for this purpose, e.g., displaying their attentiveness by imitating natural gaze patterns, facial expressions—or providing linguistic and nonverbal backchannel feedback (Skantze et al., 2015).

A motivation that is related and mentioned often in research papers is to make the interaction (appear to be) more responsive and more efficient (Ward, 1996) or to increase the perceived "fluidity" (Cassell and Thórisson, 1999). More specifically, the ability of an agent to provide appropriate backchannels has often been motivated by the increased sense of *rapport* that could be achieved (Gratch et al., 2006, 2007). Rapport describes how close to each other participants feel during an interaction and can be broken down conceptually into mutual attentiveness, positivity, and coordination (Tickle-Degnen and Rosenthal, 1990). Ward and DeVault (2016) pointed out rapport as one of the benefits of systems that take user emotions and feedback into account, alongside naturalness and improved task performance.

In other research papers, the primary motivation to endow agents with feedback capabilities goes beyond these considerations that focus on low-level functions and instead they see agent-feedback as foundational to the collaborative nature of dialogue (Brennan and Hulteen, 1995) and its importance in the process of constructing common ground (Jonsdottir et al., 2007; Kopp et al., 2007, 2008).

A completely different line of motivation—that many papers mention—is that providing feedback to users increases the perceived "naturalness" (Al Moubayed et al., 2009), "human-likeness" (Edlund et al., 2008), "life-likeness" (Cassell and Thórisson, 1999), and "credibility in behavior" (Bevacqua et al., 2008) of conversational agents. The hope is that conversational agents providing feedback result in more pleasant (Ward, 1996) and more engaging user-experiences, and in an increase in acceptance (Cathcart et al., 2003).

From this analysis of motivations, two overarching—not necessarily independent—goals in modeling feedback production in conversational agents can be identified: (1) ensuring common ground, and (2) increasing the perceived naturalness of an agent's behavior. These goals can be pursued in different ways. In order to generate agent feedback, some work models the dialogue and language processing aspects underlying feedback. In contrast to this, other work is primarily concerned with creating natural and believable agent behavior. We call these approaches **grounding-focussed** and **surface-focussed**, respectively.

3.2. Feedback and Error Handling in Spoken Dialogue Systems

In the early days of spoken dialogue systems, ASR errors were very frequent, and so it was important to develop strategies for preventing, detecting, and repairing such errors (Bohus, 2007; Skantze, 2007). This was often done through either implicit or explicit verification (or confirmation), corresponding to the notions of *display of understanding* and *clarification request*, as discussed and exemplified in Sections 2.2 and 2.3 above. Larsson

(2003) showed how these different strategies can be mapped to the levels of understanding proposed by Clark (1996), as described in Section 2.1. Typically, the ASR confidence score was used to determine what kind of strategy was appropriate (i.e., if the score was low, a more explicit verification was used). This verification (a form of feedback) would then give the user a chance to correct any potential misunderstandings in the next turn (Skantze, 2007). Since these corrections were in themselves also associated with uncertainties, statistical models were developed for tracking the system's belief in the user's intentions over multiple turns (Bohus and Rudnicky, 2005), eventually leading to a substantial body of work in what is called *Dialog State Tracking* (Williams et al., 2014) and the use of reinforcement learning to learn optimal strategies for resolving uncertainties (Rieser and Lemon, 2011). Using the distinctions made above, this line of work can be described as grounding-focused.

In parallel to this, much research has been done on how conversational agents should be able to provide verbal-vocal and multimodal backchannel-like feedback. However, this type of feedback in conversational agents is usually produced in a way that is much less deliberately planned, often in parallel to regular conversational actions (and by parallel system components), i.e., in a more surfaced-focused manner. Feedback of this type often does not appear in the dialogue record and is not generated in order to track the level of mutual understanding between the agent and the human interlocutor.

3.3. Detecting Backchannel Feedback Opportunities

A central task for agents that generate feedback in response to their human interaction partners' utterances is to determine when to respond with a feedback signal. As discussed in Section 2.5, speakers use certain cues to invite feedback, and researchers have looked into various ways of detecting these cues. Koiso et al. (1998) learned a decision tree using prosodic and a simple syntactic feature (part-of-speech, POS) to predict opportunities for producing feedback. Similarly, Cathcart et al. (2003) combined pause-durations with an n-gram part-of-speech model. Meena et al. (2014) present an online backchannel-prediction system that uses lexico-syntactic and prosodic features extracted in real-time using automatic speech recognition.

A comprehensive analysis of mostly paraverbal backchannel "inviting" cues in task-oriented communication of speakers of American English is presented in Gravano and Hirschberg (2011). This work analyses speakers' intonation, intensity, pitch, IPU duration, voice quality, as well as part-of-speech bigrams and shows that the likelihood of a backchannel happening increases quadratically with the number of cues that a speaker displays at any moment. Gravano and Hirschberg (2011) also show that speakers differ in the way they use cues. One speaker in their corpus relied on only two features (intonation and POS-bigrams) for producing backchannel cues, while most speakers used four or five, and some even all six features that were found in the analysis.

This indicates that there are individual differences in backchannel elicitation, but also raises the important questions of whether it is really the speaker using idiosyncratic cues or rather the listener responding only to some of the cues. One problem of the corpus studies of feedback behavior discussed above is that backchannels are (often) optional. Backchannel-inviting cues are, however, identified by looking at the backchannels that are actually present in a corpus and then analysing the speaker's behavior immediately preceding a backchannel. The result is that this approach does not allow the identification of backchannel cues that were not responded to. Thus, the speaker in Gravano and Hirschberg (2011) who was thought to only use two features for their cues might well have produced cues that consisted of more than two features, but to which their dialogue partner did not respond with backchannels.

This problem was addressed in the MultiLis corpus (de Kok and Heylen, 2011), which was collected in a study where three listeners were made to believe that they were in one-on-one dialogues with a speaker who was in fact talking to only one of them. The corpus thus contains backchannel responses from multiple listeners and made it possible to analyse how different listeners react to a speaker's behavior. Analyses showed that there are places in the speech where only one or two listeners respond with a backchannel but that there are also places where all three listeners responded (de Kok, 2013). This made it possible to detect more places where the speaker might have produced a cue and also shows that some cues might be more prominent than others and thus have a higher probability to elicit a backchannel response.

A similar approach to collecting multiple listener responses for the same stretch of speech uses a method called "parasocial consensus sampling" (Huang et al., 2010b), in which participants have a "parasocial interaction" (i.e., they pretend to be in interaction) with a video of a speaker. Here multiple participants were asked to respond to the speaker by simply pressing a button whenever they felt that providing listener feedback would be appropriate. It was also shown that this data collection method can be used reliably using crowdsourcing (Huang and Gratch, 2012). Heldner et al. (2013) used a parasocial interaction approach to collect richer behavioral responses by letting participants produce verbal and nonverbal feedback.

Having responses from multiple listeners enables the development of more accurate models for backchannel-prediction (e.g., based on probabilistic sequential models) that do not simply trigger a backchannel based on a rule but continuously emit probabilities resulting in a smooth probability curve that should exhibit regions of high probability during backchannel relevance spaces and can be used for feedback generation by finding probability peaks (Huang et al., 2010a; Morency et al., 2010; de Kok, 2013).

Recently, neural networks (Mueller et al., 2015) and deep-learning methods based on neural language models are being used for backchannel timing prediction. Ruede et al. (2019) combined acoustic features (pitch and energy) with linguistic features in the form of word embeddings to train an LSTM-model and could show that using linguistic features encoded in this way is useful. Using reinforcement-learning and taking the resulting level of engagement as a reward signal, Hussain et al.

(2019) learned a deep Q-network for backchannel-generation in human-robot interaction.

Most of the work for determining backchannel timing described in this section subscribes to the view that backchannels are elicited (or invited) by the speaker, i.e., that speakers (implicitly or explicitly) mark a backchannel relevance space using behavioral elicitation cues. When listeners detect such cues, they may then respond by producing a backchannel. This concept of feedback is often adopted in surface-focussed approaches to backchannel generation as natural backchanneling behavior can be produced based on rather shallow analyses of speaker behaviors.

3.4. Generating/Selecting an Appropriate Function

Agents that generate human-like feedback need not only to be able to decide *when* to produce a signal, but also *how* to express it, i.e., a specific signal that is appropriate for the given dialogue situation needs to be chosen. More specifically, it needs to be decided which communicative function (contact, perception, understanding, …; see Section 2.1) and polarity (positive, negative) should be expressed, and how function and polarity can be expressed with concrete agent behaviors and/or conversational actions. General considerations that can be found in feedback generation models are *perceptual*, *affective*, and *cognitive* aspects of the interaction (Cassell and Thórisson, 1999; Kopp et al., 2008; Wang et al., 2011).

Agents that are modeled in a surfaced-focussed way often do not explicitly choose a specific feedback function to express. When predicting that a backchannel should be given based on perceptual features of the interaction partner's behavior (see Section 3.3), these agents reactively produce a behavior such as a short verbal-vocal backchannel, a head nod, or a multimodal behavior—possibly sampling from a distribution of relevant behaviors (Gratch et al., 2006; Morency et al., 2010; Poppe et al., 2013). To produce these concrete behaviors an agent may additionally need to take its simultaneously ongoing behaviors into account. If the face-modality, for example, is already in use with a higher priority, feedback cannot be expressed with a facial expression and a feedback signal that does not rely on face-cues needs to be chosen instead (Bevacqua, 2009).

A major limitation of surface-focussed approaches for predicting feedback placement and choosing feedback form is that they lack an interactional need to produce a certain feedback signal at a certain point in time. For these systems, producing listener feedback is not a means to an end, but an end in itself. The perspective that these systems have on feedback is that it is a (surface) behavior that is desirable for agents to exhibit in order to show natural (i.e., human) listening behavior and facilitate the interaction by making human speakers believe that they are listened to and encouraging them to continue speaking (which is a function of feedback, cf. Goodwin, 1986). While making humans believe that they are being listened to can be useful for an interactive system (e.g., in order to project confidence in its ability to generate a response even when there is a processing delay that causes a long gap in turn-taking, or to create rapport between user and system Gratch et al., 2007), a shallow approach to feedback generation may actually thwart the intended effect. If the agent's feedback suggests its ability to understand what the human interaction partner says but then fails to respond in a meaningful way, confidence in the system will likely vanish.

Agents that are modeled in a grounding-focussed way aim to generate more nuanced feedback behavior, that is grounded in cognitive or affective states, take a variety of approaches to the function/form selection process. For these agents, the ability to detect backchannel relevance spaces can be considered a necessary—but not a sufficient—ability to produce feedback.

3.4.1. Affective Considerations

Agents that take affective considerations into account for feedback generation may generate a broad range of feedback functions that are meant to convey their (simulated) affective or emotional state, display similarity, or influence the user's emotional state. For Cassell and Thórisson (1999), this means that an agent is able to generate emotional feedback with functions such as agreement (resulting in a smile) or confusion (resulting in a facial expressions that conveys puzzlement). The SimSensei agent (DeVault et al., 2014), a virtual interviewer in a mental healthcare domain, is able to provide affective feedback with the function of being empathetic to or surprised about what the user says. Prepin et al. (2013) describe a model that can express the dynamically evolving "dyadic stance" by adapting smile-behavior to the interaction partner, in a way that displays either mutuality or divergence.

Affective considerations in choosing feedback functions can also go beyond transient affective states. The SEMAINE-project (Schröder et al., 2012), for example, explored the influence of personality and emotion and developed a set of four "sensitive artificial listeners," embodied conversational agents—each with a different simulated personality (aggressive, cheerful, gloomy, or pragmatic)—able to generate multimodal feedback in response to a human speaker's conversational behavior. Feedback planning in this model is based on acoustic, visual (head nods and shakes) and simple linguistic features (words) extracted in real-time from the human interaction partner's behavior and analyzed for their affective content. In evaluation studies it was shown that users preferred the affect-sensitive listener to a control and that it led to higher engagement, felt engagement, and flow.

3.4.2. Cognitive Considerations

Kopp et al. (2008) proposed a model for a feedback system of an embodied conversational agent that incrementally considers the (typed) utterance of the human interaction partner and evaluates it on the the basic levels of communication that feedback serves according to Allwood et al. (1992), i.e., contact, perception, understanding, acceptance, emotion, and attitude (see Section 2.1). To evaluate the agent's ability to perceive, it is checked whether incoming words are in the agent's lexicon, to evaluate its understanding, the agent checks whether it can interpret the input. The agent's acceptance is evaluated by comparing the intention of the utterance to the agent's own beliefs, desires, and intentions. Based on these evaluations, the system then generates a functionally appropriate feedback signal at a specific point

in time (and with a specific form), combining two planning mechanisms: Reactive behaviors are triggered by a set of rules that respond to events in input processing, whereas more deliberate behaviors are generated in a probabilistic fashion based on the development of the agent's longer-term **listening state**.

Similarly, the model of Wang et al. (2011) incrementally processes user utterances and computes a partial semantic representation of it. Based on confidence values associated with its understanding of the partial utterance, the model computes a score and positions itself in one of three states (confusion, partial understanding, understanding) that are then used to generate specific feedback signals.

3.4.3. Further Considerations

Wang et al. (2013) further introduced aspects of *conversation roles* into feedback generation in conversational agents. Depending on the role of the agent (addressee, side-participant, overhearer) as well as its goal (participation goal, comprehension goal) their model chooses different functions and forms of feedback, e.g., in order to express its roles or to signal an intended change in its role.

3.5. Generating Different Forms of Feedback

After choices of feedback timing and function have been made, the last aspect that needs to be determined for feedback generation is the specific form of the feedback signal that should be generated. As shown in Section 2.4, backchannels take the form of verbal-vocal signals that can be varied phonologically, morphologically, and using prosody (Allwood, 1988; Ward, 2006). In addition to this, backchannels can be nonverbal or multimodal in their form when displayed through head gestures, facial expressions, eye gaze, etc. (cf. e.g., Allwood et al., 2007; Włodarczak et al., 2012). Thus, the question underlying research on the choice of feedback form is not only which form to choose to express a certain function and meaning, but crucially also the more basic question of which meaning is expressed by a specific form that an agent is able to produce.

The expression of nuanced differences in meaning and function through prosodic variation of verbal-vocal feedback signals has been explored by Stocksmeier et al. (2007), who synthesized 12 prosodic variants of the German backchannel "ja" (*yeah*) using different intonation contours and had them rated on semantic differentials such as for example hesitant–certain, happy–sad, approving–rejecting. They found that agreeing/happiness, boredom and hesitancy were attitudes that could be distinguished most clearly for certain variant clusters. Similarly, Edlund et al. (2005) and Skantze et al. (2006) changed the focal accent peak of fragmentary grounding utterances ("red?") and could show that they are perceived to convey different grounding categories (accept, clarify understanding, clarify perception) and that they influenced participants' subsequent responses in a human-agent interaction study.

In contrast to these prosodic variations, Kawahara et al. (2016) varied the morphological form of Japanese backchannels through repetition/reduplication ("un," "un un," "un un un"; cf. Allwood, 1988) and showed that the variant to use can be predicted by the boundary type of the preceding clause and by the syntactic complexity of the preceding utterance (number of phrases).

Oertel et al. (2016) devised perception test methods to investigate the perceived level of attentiveness that is conveyed by multimodal feedback expressions of a virtual listener, such as prosodic variation of verbal or vocal form, as well as head nods. Using this method they were able to identify features that led to an increase of the perceived level of attentiveness.

A model that can produce a number of different feedback forms is described by Bevacqua (2009). It defines a lexicon of multimodal feedback signals that an agent can produce while interacting with a user. For the model it was studied (details in Bevacqua et al., 2007; Bevacqua, 2013) which meaning humans assign to these signals, finding that positive and negative feedback can be reliably separated but that signals can be both polysemous and synonymous regarding more specific functions—i.e., users may assign different meanings to the same signal and the same meaning to multiple signals. For each meaning the authors are, however, able to identify signals that are relevant for application in the sense that they are recognized as such by a majority of their participants. The final feedback lexicon of Bevacqua (2009) consists of a set of rules to express feedback functions. Each rule captures core behaviors that define a signal as well as additional behaviors (on other modalities) that can be added in some situations. Furthermore, the model also takes into account which modality is currently available for producing feedback.

One of the foundational ideas of embodied conversational agents, in contrast to voice-only conversational agents, was to be able to express interactional functions, such as feedback, using bodily means (Cassell, 2001). In order to synthesize such functions in multimodal behavior, it is important to be able to coordinate the behaviors on individual modalities (and with the user) with regard to timing and resources. A successful general solution for this problem was developed in the SAIBA-framework with the standardization of description-formats for multimodal behavior (the "Behavior Markup Language," BML; Kopp et al., 2006; Vilhjálmsson et al., 2007) that can serve as specifications for synthesis in behavior realization engines (Kopp and Wachsmuth, 2004; Thiebaux et al., 2008; van Welbergen et al., 2009). Using this basis, multimodal agent-feedback behaviors could be described on different levels of abstraction (e.g., provide positive feedback, synthesize a head nod with a verbal positive feedback, synthesize head nod variant X while gazing at object Y and uttering "uh-huh" 300 ms later) and using different synthesis methods (e.g., parametric, key-frame-based, or motion-captured animations). A significant number of embodied conversational agents able to provide multimodal feedback were built upon systems that use this framework (e.g., Kopp et al., 2008; Morency et al., 2010; de Kok and Heylen, 2011; Wang et al., 2011; Schröder et al., 2012).

Embodied means of providing feedback spans different levels of awareness, control and intentionality (Allwood et al., 2007). Blushing, for example, is a feedback behavior indicative of an inner (bodily) state of which listeners are not necessarily aware, have limited control over, and do not communicate intentionally. A head nod, on the other hand, is a feedback signal that is intentionally communicated, similarly to

linguistic communicative acts. Other feedback behaviors may lie in between.

In general, the form of a feedback behavior has an influence on its perceived meaning and function and interacts with its dialogue context (e.g., Allwood et al., 2007). A systematic study of different nonverbal feedback behaviors does not exist, yet (but see Bevacqua, 2013 for a survey). The "head nod," however, as the most prototypical nonverbal feedback behavior, has received a considerable amount of attention from different fields (Hadar et al., 1985; Cerrato, 2005; Heylen, 2008; Petukhova and Bunt, 2009; Poggi et al., 2010; Włodarczak et al., 2012; Ishi et al., 2014). In the field of conversational agents, head gesture generation and evaluation, has, however, mainly focussed on head movements (including nods) as co-speech speaker—in contrast to listener—behavior (e.g., Lee and Marsella, 2010; Ding et al., 2013). Head nods are part of the behavioral repertoire of most listening agents, but detailed analyses of how a listener agent's head gestures are perceived by users are rather sparse. Bevacqua et al. (2007) let participants assign backchannel-meanings in three categories (performative, epistemic, affective) to short videos of a virtual agent producing different head gestures and facial expressions (also in combination)—the results of this study were already described above. Oertel et al. (2016) synthesized head nods from data and had them rated for perceived attentiveness of an agent. The result shows that for a head nod to communicate attentiveness, it should be rather long, have multiple oscillations and be overall more energetic.

3.6. Implications for the Future Design of Conversational Agents

The work we reviewed in this section shows that different aspects need to be considered when developing conversational agents that should be able to provide conversational feedback in interaction with users. While the timing of feedback signals has received a considerable amount of attention from the research community, resulting in well-performing technical models for determining backchannel relevance spaces, the choice of function and form of feedback signals are less well understood, yet. This may be acceptable for conversational agents that operate in scenarios where a surface-oriented approach to feedback generation is sufficient—e.g., for entertainment purposes or in settings where the goal is to simply encourage users to keep talking. Future conversational agents, that are more conversationally competent (i.e., able to more deeply and broadly understand what their users mean and able to detect and repair trouble and miscommunication in the interaction), need to aim for a more grounding-oriented approach toward feedback generation. For them to be able to more deliberately use feedback (deciding which feedback function to produce in order to advance the interaction), more research is needed on the integration of feedback generation models with models for language understanding, dialogue state, and dialogue management. To then be able to synthetically produce a (multimodal) feedback signal that communicates the intended function and meaning, more research is needed to better understand the interaction between the large multidimensional space of the form of multimodal feedback signals and their multifunctional meanings. The insights gained from these two research directions will also be highly relevant for creating conversational agents that can deal with user feedback, as described in the following section.

4. FEEDBACK FROM USERS TOWARD AGENTS

This section presents work on how a conversational agent can identify, analyse, and understand multimodal feedback. Both recent and more ground-laying work is presented. Of special interest is work that classifies feedback either in terms of the grounding levels described in Section 2, or in terms of the user's attitudinal reaction to the system, mentioned in Section 2.1.

To interpret feedback provided from a user toward a system, the system must identify the signals (Section 4.1) and map them to some representation (Section 4.2). If the user does not produce signals understood by the system, the system can in turn elicit feedback from the user (Section 4.3). We conclude this chapter by summarizing the implications for the future design of conversational agents (Section 4.4).

4.1. Identifying Users' Feedback Signals

As described in Section 2, feedback signals can serve multiple purposes and may be delivered in many different ways across many different modalities. It follows that different types of sensors are appropriate for different types of scenarios to let agents sense the feedback of their human interaction partners. Identifying a signal is a precondition for a system to understand it, just like in the human-human principle of *upward completion* described in Section 2.1, where identification is a precondition for positive or negative understanding.

A parallel problem to that of identifying feedback is identifying what is *not* feedback, or what is *incidentally* feedback (what Allwood et al., 1992 call "indicated"). A feedback signal can be unintentional but still carry information that is relevant to the context. One way of identifying such signals, at least in annotation, is by using *salience* as a criterion (Brunner and Diemer, 2021).

4.1.1. Speech

In Section 2.4, Clark's model with two *tracks* on which communication can happen (Clark, 1996) was contrasted to Yngve's main and back channel model (Yngve, 1970). Speech is the modality where the distinction between these two feedback perspectives is the most explicit, since main channel contributions and *track 1* contributions are typically speech.

A system that is aware of backchannel signals will want to separate those signals from main channel contributions. This separation is crucial for being able to analyse the feedback in terms of grounding. Gustafson and Neiberg (2010) and Heldner et al. (2010) have shown that the pitch of a backchannel is typically similar to the pitch of the speech being responded to. While pitch alignment of this type also happens for main-channel contributions, Heldner et al. (2010) show that it is more common for backchannelling. Additionally, the data

analyzed by Gustafson and Neiberg came from a corpus where the conversations took place over telephone, indicating that the prosodic adaptation patterns still applied even with the slight latency of a telephone conversation compared to in-person speech.

Skantze et al. (2014) performed an experiment where a robotic instructor guided a human participant through a map. They showed that the performance of participants in the map task was connected to the user's utterance timing and prosody, and that this was more important than the linguistic contents of their speech. This is an example of how the pitch of speech, sensed through rising or falling F0, can be used to sense the tone of a user's utterance.

The verbal, linguistic component of speech can be sensed through automatic speech recognition. The quality of speech and confidence of the ASR can be used to inform the system's behavior. Early work in this field was performed by Brennan and Hulteen (1995), who extended Clark's feedback ladder (see Section 2.2) with more explicit levels suited for a system conversing over a phone line. Through a Wizard of Oz experiment, Brennan and Hulteen showed interesting grounding criterion properties of this scenario. For example, users were more accepting of requests for clarification and displays of clarification directly following a corrected mis-idenfication. Modern smart speakers are also speech-only devices, and thus share their only communication modality with the phone-only platforms of these early systems.

Dialogue acts are a way to annotate speech in a dialogue by its communicative function. They are a generalization of *speech acts*; where speech acts annotate the communicative function of a single utterance by a speaker (Searle, 1969), dialogue acts instead annotate what function utterances have in the context of the dialogue (Allen and Core, 1997; Core and Allen, 1997; Jurafsky et al., 1998). Feedback falls under what Core and Allen (1997) call *Backward Communicative Functions*, which are the dialogue acts that refer to previous utterances in the dialogue. Thus, properly classifying speech as dialogue acts also means identifying whether it is feedback or not, but crucially, since grounding models like those by Allwood et al. (1992) or Clark (1996) are not part of the dialogue act tag standard, annotating speech acts does not say what the feedback means, only that it is feedback. In Section 5.3, some annotation schemes that extend dialogue acts with representations of grounding are presented.

Shriberg et al. (1998) used prosody and ASR to classify speech as dialogue acts. Hanna and Richards (2019) have shown that a clear understanding of a conversational system's intended dialogue act correlates with high acceptance and understanding of the system's proposals. A similar high-level approach for tagging user speech in modern ASR systems is *intent classification*, where speech of some length, typically sentence-length or longer, is classified to approximate the *intention* (Ajzen, 1991) of the user (Purohit et al., 2015; Larson et al., 2019).

Intent classification typically assumes that the user's intention can be found by analysing their utterance as text—the work by Purohit et al. (2015), who applied intent classification to short messages on Twitter, showed that it is possible to extract intent when the channel is text-only and restricted in length. When the main modality is speech, supported by possible side modalities like gaze and gesture, however, it is not certain that the user's full intention is actually captured by a transcript of their spoken utterance.

More recently, dialogue act classification has moved on to using more sophisticated machine learning approaches, often using the same Switchboard corpus as Shriberg et al. (1998). An example of this is the approach by Liu et al. (2017), who showed that convolutional neural networks can classify dialogue acts on the Switchboard corpus relatively well, and that adding meta-information about speaker shifts and what dialogue acts preceded the given context increases the classification accuracy. Qin et al. (2020) recently presented work showing that state-of-the-art dialogue act classification can be obtained by classifying the dialogue act in parallel with the sentiment of the speech being classified. This shows that the two types of classification overlap.

4.1.2. Gaze

Nakano et al. (2003) presented a study of a stuffed toy robot with eye-gaze capacity. In this study, it was shown that eye contact between the user and the robot led to favorable feelings from the user toward both the robot and the interaction, and that mutual attention, gaze on the same object, did not lead to the same improvement when used on its own. However, the highest favorable reaction from the system came from the use of both eye-contact and mutual attention. Nakano et al. (2003) took this to mean that the main effect of mutual attention through gaze is subconscious, while eye contact is more easily consciously picked up by users.

Skantze et al. (2014) show that gaze can be used as a measure of uncertainty based on how it is used together with other feedback signals. The scenario used is a map task where a robot head presents a path through a map to a human user. The authors show that gazing on the robot has a positive correlation with certainty, i.e., that users look more at the robot when they are certain of the path they have just taken through the map.

Mutlu et al. (2009) find that gaze is an efficient way for a social robot to inform users of what their role in the conversation is—in cooperation with those users, as a joint act—and also find that the resulting turn-switching between users and the robot makes them feel more invested in the interaction.

4.1.3. Head Gestures

Hee et al. (2017) found that human listeners are more likely to use non-verbal feedback, including head gestures, toward embodied agents than toward non-embodied agents. An early attempt to classify head nods and head shakes as feedback in users of a conversational agent came from Morency et al. (2005). In their study, they found that classification of nods and shakes was possible to do with high accuracy when the visual data from a camera sensor was combined with a prediction from dialogue context. For the prediction, high-level features were used, such as whether the robot's previous line had started with "do you" or ended with a question mark.

Paggio et al. (2017) attempted to classify the presence of any head gesture (from frame-to-frame movement features) in a video corpus but were only able to achieve a 68% accuracy.

In a subsequent study, Paggio et al. (2020) added a binary feature to the model denoting whether the target was speaking or not, slightly raising the accuracy to 73%. They were also able to make a distinction between nods, head-shakes, and miscellaneous gestures with an F-score of around 0.4. This illustrates the difficulty of classifying three-dimensional head movement features defined both by position and rotation from two-dimensional video.

4.1.4. Body Pose and Facial Expressions

The user's body pose can be used to estimate their intent to *engage* with an embodied agent (Bohus and Horvitz, 2009; Sanghvi et al., 2011; Schwarz et al., 2014). Bohus and Horvitz (2009) built machine-learning models for predicting the measure of engagement through many features describing the user's pose—how they were standing, whether they were facing the embodied agent, and how they had moved in the recent past with several different time-scales. Bohus and Horvitz found that by using a subset of these features, systems could learn to predict the moment of engagement or disengagement by as much as 4 seconds, implying that the body pose features that humans use to display an intent to engage or disengage happen over a surprisingly long span. Engagement in this sense is a parallel concept, not quite the same as Clark's (Clark, 1996) *attention*, since a listener may want to end an interaction (disengagement) while still accepting the proposals of the speaker in the short term.

Body pose can overlap with gaze direction and facial expressions when a system wants to sense head gestures like nodding (Sidner et al., 2006) or head-shakes (Morency et al., 2005, 2007). Facial expressions (Lisetti and Rumelhart, 1998; Lisetti and Schiano, 2000) can also be used to enhance the system's understanding of speech (Kleckova et al., 2005), to sense the intention to interact (*engagement*) (Chiba et al., 2017), or to approximate the user's affective state (Khosla et al., 2012; Tzirakis et al., 2017). Engagement as a measure of whether the user wants to interact is a low level of grounding on its own, corresponding to *attention* in Section 2, but nods as a display of understanding (Section 2.2) convey understanding or acceptance, indicating a higher level of grounding.

Heylen et al. (2007) found that there were strong connections between facial expressions and feedback acts linked to positive or negative feedback on the scale by Allwood et al. (1992) (see Section 2.1), but also found that some signals were only significant in combination with other head pose and facial expressions (Bevacqua et al., 2007; Heylen et al., 2007). For example, tilting one's head was interpreted as a sign of negative acceptance if the subject also frowned. These results came from an experiment where an agent produced the facial expressions, so it is unclear how important these findings are for measuring the importance of facial expressions from users toward agents; the system could have been producing signals that were less ambiguous than those produced by the average human.

4.1.5. Multimodality

Kopp et al. (2007) have stressed the importance of multimodal incremental sensing of feedback, i.e., both sensing signals in real-time and sensing multiple signals at the same time. The previously cited work in Section 4.1 shows that many different modalities have been possible to sense for a relatively long time, and that combining features often leads to an improvement in picking up signals accurately. However, multimodal systems that can sense many different types of signals at the same time are more rare than single-modality systems, and hard to compare even when they do pick up multiple signals. Poria et al. (2017) state why this is not as simple as simply adding in more modalities to get improved performance:

> [P]oor analysis of a modality can worsen the multimodal system's performance, while an inefficient fusion can ruin the multimodal system's stability.

Chiba et al. (2016) used acoustic features in combination with facial landmarks to estimate users' mental states, although the work seemed to work best for detecting the intent to engage (Chiba et al., 2017).

Baur et al. (2016) presented a model of how a conversational agent could recognise the user's social attitude toward it. Two systems were proposed: one entirely virtual, and the other embodied by a humanoid robot. Both of the proposed systems were equipped with sensors for head tracking, ASR and eye sensors, and used Bayesian networks to map the user's behavior to an approximation of their engagement, as well as which behaviors the system should employ in response. The discourse, context, and robot's personality, as well as entrainment on an individual user, are allowed to control how likely one behavior is to lead to the activation of another. The authors do not, however, present any experimental results backing up the performance of their models.

In a survey on multimodal approaches for emotion detection, Marechal et al. (2019) show that affective state can be extracted both from the user's pose (Zacharatos et al., 2014), physiological features like blood pressure (Shu et al., 2018), and from facial features (Ekman, 1993). While the physiological modality may be impractical to record for an embodied agent in the wild, facial data, and pose data could be used, in combination, by any agent that has video cameras.

Zhou et al. (2018) present a case study on the measurement of cognitive load through multimodal physiological features connected to a computer. While the output of the systems described by Zhou et al.—an estimation of cognitive load—were not in themselves interesting for the purposed described in this paper, the work presents a good example of how the types of sensors mentioned by Marechal et al. can be used in a computer interface, how they can be multimodally fused (with beneficial results presented by Zhou et al.), and how intrusive such an interface is. Specifically, the experiment performed here measured eye movements, galvanic skin response, and task-specific measurements like the length of pen strokes – all of these apply to human-agent interaction as well. To accurately measure eye movements, the authors used special head-mounted cameras.

Guntz et al. (2017) found that a combination of modalities performed the best when classifying chess-playing test participants by their skill in the game. In this study, emotions (estimated through face interpretation as described

previously) were the most highly-performing single modality, but combining facial emotions with gaze resulted in more well-performing classifiers.

Recently, Axelsson and Skantze (2022) showed that individuals interacting with a system presenting a painting to them generally used feedback in the head and speech modalities, and that feedback in the facial, body pose and gaze modalities was not important for classifying their response as positive, negative, or neutral. These results give an indication that facial expressions are not important to sense for a presenting system of this kind, but may not be applicable in scenarios where the audience is more involved with the interaction than being the audience.

4.2. Understanding User Feedback Behavior

When a feedback signal from a user has been identified, a feedback-aware conversational system must decide what the signal means. The shape that the internal representation takes is highly scenario-specific and depends on the internal representations of the task and user models. As we pointed out in Section 3, models of how systems *provide* feedback can be surface-focussed or grounding-focussed depending on whether their motivation is to use feedback because feedback improves the quality of the interaction, or because it is justified by the state of the user, respectively. For systems that pick up user feedback, there is a distinction between systems that simply identify the presence of feedback from the users and systems that attempt to understand what it is referring to and how—but both of these are grounding approaches as described in Section 2.2, since even identifying a backchannel is a sign of attention.

4.2.1. Understanding User Feedback in Terms of Attitudinal Reactions and Affective State

Affect is a general term for the emotions felt by the user of a system. As was touched on in Section 4.1.5, it is possible to see affect as an output estimated on other, more concrete multimodal signals produced by a user. Indeed, for some multimodal systems, the main purpose is to estimate the user's affective state (Poria et al., 2017).

Tzirakis et al. (2017, 2021) present an end-to-end deep learning system which approximates the user's affective state based on multimodal input features, specifically facial images, and speech. Comas et al. (2020) estimate affective state by combining facial images and physiological features, specifically EEG and skin conductivity.

Skin conductivity and heart rate were used to estimate affective state in a human-robot interaction setting by Kulic and Croft (2007), who also argued that affective state can be used as an input feature for telling a system whether the user is expected to give feedback or not. More recently, Schodde et al. (2017) presented specific multimodal signals together with a model for what they mean in terms of children's affective states when interacting with a social robot.

4.2.2. Understanding User Feedback in Terms of Dialogue Acts

In more recent end-to-end approaches for dialogue modeling, speech, and feedback recognition can be seen as a classification task from user behavior to the state of the system (Qian et al., 2017; Shi et al., 2017), or to the value of specific slots in a slot-filling dialogue state (Xu and Hu, 2018; Ma et al., 2020; Ouyang et al., 2020; Zhang et al., 2020). A generalization of this, that moves beyond task-specific dialogue systems, is to classify speech and feedback as dialogue acts instead (Liu and Lane, 2017; Ortega and Vu, 2017). End-to-end models can be trained to contain internal states keeping track of the user model, learning these states simply from large amounts of dialogue examples (Crook and Marin, 2017; Boyd et al., 2020; Li et al., 2020). A restriction of these large-scale models is that they typically only work on text-to-text data and do not extend to multimodal input or output.

4.2.3. Understanding User Feedback in Terms of Grounding

Axelsson and Skantze (2020) use knowledge graphs based on Wikidata to map user feedback to structured information that the system can use to decide how to express its dialogue. The same concept is used independently, in an extended manner by Pichl et al. (2020), who insert objects into the knowledge graph representing the user and the system, and connect those objects to WikiData objects through relations expressed during the dialogue. **Figure 2** illustrates the approach taken in Axelsson and Skantze (2020): Feedback from the user, which can be either verbal or non-verbal, marks edges of the knowledge graph as grounded or ungrounded on the four levels defined by Clark (1996) (see Section 2.2). Individual nodes of the graph can be marked as more or less known by an individual, causing our dialogue system to refer to the entity by shorter references, or pronouns if appropriate. The approach works well for mapping a presentation task, where the robot has the initiative and the user's responsibility is to react with feedback, even when there are multiple users. The approach by Pichl et al. (2020) may be more appropriate for less presentation-oriented dialogues.

A different approach is taken by Buschmeier and Kopp (2014), who model feedback interpretation and representation in terms of an "attributed listener state" (ALS). In this model of feedback understanding, the user's feedback behaviors, relevant features of the agent's utterances, and the dialogue context are used in a Bayesian network to reason about the user's likely *mental state of listening*, more specifically whether contact, perception, understanding, acceptance and agreement (see Section 2) are believed to be low, medium, or high (Buschmeier and Kopp, 2012, 2014; Buschmeier, 2018). As **Figure 3** illustrates, this inference can be done incrementally while the agent is speaking—e.g., at every backchannel relevance place—so that the agent always has an up-to date idea of how well the user is following its presentation. Buschmeier (2018) calls this process a "minimal" form of mentalizing (following the concept of a "most minimal partner model" Galati and Brennan, 2010), which enables the agent to adapt its presentation in a high-level fashion, e.g., by

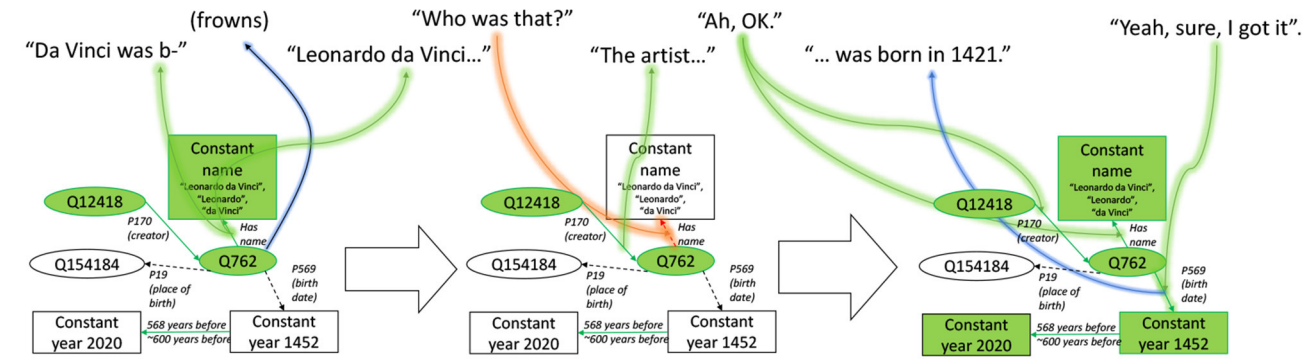

FIGURE 2 | An example of the knowledge graph approach for grounding mentioned in Section 4.2.3, adapted from Axelsson and Skantze (2020).

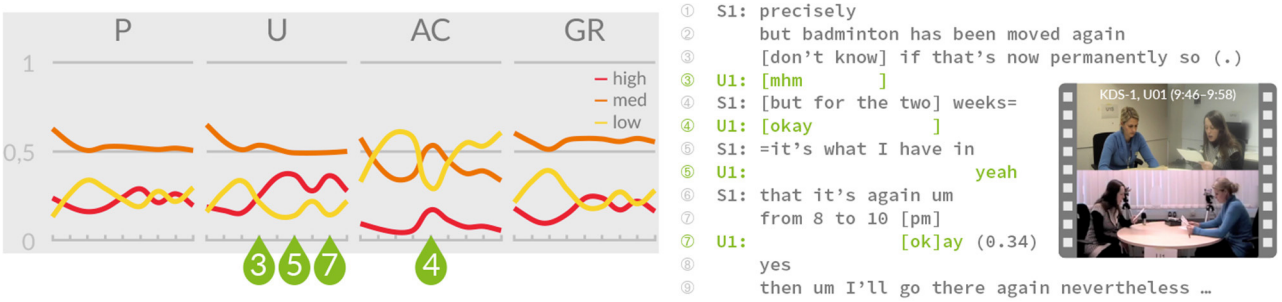

FIGURE 3 | In the attributed listener state model of feedback understanding, the listener's likelihood of perception (P), understanding (U), acceptance (AC), and grounding-potential (GR) is inferred incrementally while the speaker is presenting information (adapted from Buschmeier and Kopp, 2014).

being more redundant or by repeating information that is already considered grounded.

For systems that only handle one domain or type of task, like the calendar agent used as an example by Buschmeier, this minimal mentalizing approach might already be sufficient. For systems that present more arbitrary information, like the poster presenting system by Axelsson and Skantze (2019), negative understanding toward one part of the utterance may imply positive understanding toward another part, or perhaps invalidate the system's previous belief that the user understood something earlier, and thus feedback must also be interpreted and handled in terms of some kind of dialogue state or more complex partner model, which becomes similar to a representation of full common ground—with higher associated costs for language production and adaptation (Keysar, 1997).

4.3. Eliciting User Feedback

A system that can handle multimodal user feedback needs to be able to handle the case that the user does not give any feedback, or that the user does not give enough feedback for the system to create an effective user model. Feedback could be missing because the user has not given any feedback, or because it was given through signals that the system is not equipped to sense. One way to address missing feedback is to **elicit** feedback, or elicit feedback in specific modalities through cues. In human-human interaction, elicitation cues can be both prosodic and gesture-based (Bavelas et al., 1992, 1995; McClave, 2000), or syntactically related to what the speaker says (Gravano et al., 2012). Buschmeier (2018) presents this as a way for a speaker to address a mismatch in **information needs**, i.e., when the listener does not believe that the speaker needs feedback to keep talking, but the speaker does. The types of elicitation that apply for agent-human interaction were discussed in Section 3.3.

When elicitation cues are used by conversational agents, it has been shown that the cues are actually connected to responses by the human conversation partner, indicating that elicitation cues can be employed by conversational systems (Misu et al., 2011a,b; Reidsma et al., 2011; Buschmeier, 2018). Gaze cues from an agent toward the user can be an effective way to cue the user to display proof of attention (Frischen et al., 2007). Hjalmarsson and Oertel (2012) showed that the gaze behavior of an on-screen agent could be used to elicit backchannels from a user, with more backchannels appearing when the agent looked at the user more. This was elaborated upon by Skantze et al. (2014), who showed that a physical robot with a face could use its gaze to elicit feedback by either looking at the task (a map game) or the user.

4.4. Implications for the Future Design of Conversational Agents

As we argued in Section 4.1.1, dialogue acts and intent classification do not contain enough information to classify whether a user's reaction to a dialogue system means positive or

negative grounding. A scheme that extends dialogue acts to cover this are *grounding acts* (Traum and Hinkelmann, 1992; Traum, 1994), which generalize dialogue acts to contextualize what the listener's responses mean in terms of the state of the conversation. Schemes like this indicate that the classification of user feedback into more feature-rich representations is a way to understand the user more deeply.

In addition to changing how feedback is classified, an important future direction for the field of understanding user feedback is integrating the production and interpretation of feedback. This merges the theory presented in Chapters 3 and 4. This argument is elaborated upon in Section 5.4.

5. FUTURE DIRECTIONS

5.1. Toward Grounding-Focussed, Continuous Agent Feedback

In Section 3, we presented the difference between *surface-focussed* and *grounding-focussed* systems, where the former produces feedback because it is believed to have a global positive effect on the interaction (such as increased rapport or sense of engagement). In the latter type of systems, choices about which feedback to give at which point in time are made based on how instrumental they are to increase the level of grounding of individual pieces of information. This type of feedback is more complex to model, since the system needs to keep track of these information units and their associated grounding status, whereas surface-focussed models do not usually "remember" what they have actually given feedback to. Note that grounding-focussed feedback can also be instrumental to increasing global interaction quality objectives, just like surface-focussed feedback.

As we have seen, there has been a lot of work in traditional dialogue system research on how to signal a system's level of understanding, in the form of clarification requests (negative feedback) or display of understanding (positive feedback), which the user then can react to, so that the system can make sure that it has understood the user's intentions in spite of uncertainties (stemming, for instance, from ASR and NLU). Thus, this tradition has had a clear grounding focus.

When it comes to more continuous feedback, in the form of backchannels, there has been a lot of research on how to produce such listener responses at appropriate places while the user is speaking (i.e., when the user provides a backchannel-inviting cue) to give the impression of an agent that is attentive (see Section 3.1). In such approaches, this feedback does not reflect the agent's actual level of understanding, and often there is no deeper processing of the user's speech. Thus, the feedback is not produced to reach mutual understanding, but rather as some kind of support for the speaker. While there might be certain use cases for such models, this approach is clearly limited. In fact, such feedback can in many cases be misleading and counter-productive, as it might give a false impression that the agent is understanding the user while it is not. When the lack of understanding is eventually revealed, users may lose trust and confidence in an agent.

We think there is a need for more research on how to integrate continuous feedback models with more traditional, grounding-focussed, models of mutual understanding (such as dialogue state tracking), so that the agent can produce listener responses that reflect a deeper understanding.

At the same time, this type of modeling is not in line with the recent trend of end-to-end modeling in conversational systems. In such approaches, dialogue data is collected and a model is trained to generate the behavior seen in the data. This can be fine when it comes to providing answers to questions, etc., but it is hard to see how it can be applied in a meaningful way to the modeling of feedback. Sometimes a listener will produce negative feedback on some level and sometimes positive feedback on some other level, but the reason why the listener produced the specific feedback was based on their current level of understanding, and this information is not overtly present in the data. Thus, an agent trained on such data could learn to produce various forms of feedback which might sound appropriate, but it would not be instrumental in reaching mutual understanding.

5.2. Timing, Form, and Function of Agent Feedback

So far, most work on the automatic generation of backchannels has focussed on the timing of backchannels (see Section 3.3). This is often motivated by surface-level objectives. However, if these backchannels should also reflect the agent's level of understanding and attitude, future work has to look more into the form and function of backchannels. For example, if the agent only commits to the level of continued attention, the prosodic realization of the backchannel should be different from cases where it wants to signal agreement. So far, research on synthesized backchannels has been very limited (Stocksmeier et al., 2007; Pammi, 2011), and off-the-shelf synthesizers typically do not have a comprehensive library of backchannels with associated "meanings," or useful parameters to control the prosody either.

This is important also for surface-focussed systems, since the production of certain forms of backchannels are likely to sound very "off," given the preceding context, even if the timing might be appropriate (Poppe et al., 2013). Such backchannels are likely to be detrimental to the user's experience of rapport and engagement.

A similar problem exists for multimodal feedback generation. Although there have been studies on the perceived function and/or "meaning" of facial expressions, head gestures, and manual gestures of artificial conversational agents, it is not clear how these findings could be mapped onto different embodiments and how functions interact when behaviors on different modalities are combined. Although questions regarding the compatibility of multimodal signals have received significant attention in standards for multimodal output generation (the SAIBA framework; c.f. Kopp et al., 2006; Vilhjálmsson et al., 2007), the focus was on low-level constraints such as timing and availability of individual modalities. Concepts regarding higher-level aspects, such as function combination and function to behavior mapping are less well understood (Heylen et al., 2008;

Cafaro et al., 2014). In general, it is unclear how planning-based multi-level multimodal output generation can be adapted to work in a dynamic and incremental human–agent interaction (Buschmeier, 2018, Section 8.3), the integration of backchannels and grounding phenoma into incremental dialogue models gained interest in recent years (Visser et al., 2014; Eshghi et al., 2015).

5.3. Interpreting Multimodal, Context-Sensitive User Feedback

Compared to other forms of communicative acts, the interpretation of feedback is perhaps the most challenging. The reason for this is that feedback often has fragmentary form with little syntax to guide the interpretation. Instead, meaning is often conveyed through prosody and visual signals, and the interpretation is often highly dependent on the context and task at hand. Attempts have been made to map user feedback to the different levels of understanding discussed in Section 2.1, but this has turned out to be very hard. This might help to explain why there is less work on handling user feedback, compared to work on how agents should be able to give feedback.

The standard method for understanding the user's speech in conversational systems (i.e., NLU) is to use some form of intent classification. However, in Section 4.1.1, we argued that dialogue acts and intent classification are not necessarily strong enough frameworks to model how a user's feedback to a system construes positive or negative grounding. While dialogue acts may not directly say anything about the user's grounding state themselves, Traum and Hinkelmann (1992) proposed an extension to dialogue acts called *grounding acts* (further developed in Traum, 1994). This scheme is a generalization of dialogue acts, describing what the listener's responses mean for the conversation. The dialogue act classification systems described in Section 4.2 show that machine learning approaches are quite viable for classifying dialogue acts; an interesting direction for future work would be to use those same approaches for classifying grounding acts instead, and to see if this enables generally grounding-aware dialogue systems. Benotti and Blackburn (2021) recently argued that grounding-aware dialogue systems are only viable if they can handle negative grounding from their users, and consider classifying grounding acts automatically to be a future direction.

Moreover, intent classification in current systems often only relies on the text (from the ASR), whereas a classifier of feedback and grounding would also need to take prosody (and visual signals) into account, which brings it closer to the area of *social signal processing* (Vinciarelli et al., 2009).

5.4. Bidirectional Feedback in Mixed-Initiative Dialogue

Most work so far has focused on either agent feedback or user feedback (as outlined in Sections 3 and 4, respectively), and depending on which one has been in focus, different domains and applications have been used. In "classic" dialogue system domains, such as ticket booking or question answering, the focus has been on how agents can provide feedback (such as clarification requests) to catch errors and misunderstandings (as discussed in Section 3.2). In studies on agent backchannels, typical domains include story telling (Schröder et al., 2012), instructions (Meena et al., 2014) or interview scenarios (Johansson et al., 2016), where the user is talking most of the time. When it comes to user feedback, domains where the agent is talking most of the time are typically selected, such as agent-human presentation scenarios (Buschmeier and Kopp, 2018; Axelsson and Skantze, 2020).

Less work has been done on integrating these different forms of feedback into one system or one model. One example of such an integrated system was the number dictation system presented by Skantze and Schlangen (2009), where the user was reading a number sequence to the system while the system provided continuous feedback, in the form of backchannels, displays of understanding and clarification requests, which it keeps track of, as well as the user's reaction to them. The system then reads back the sequence, allowing the user to give continuous feedback in a similar fashion. However, the domain itself was clearly very limited. More complex domains allowing for mixed-initiative dialogue, where bidirectional feedback is relevant, should be explored. There is some evidence suggesting that when agents provide feedback, users also expect their own feedback to be understood by the agent (Sidner et al., 2006; Kontogiorgos et al., 2021; Laban et al., 2021).

We can expect synergies to emerge when feedback generation and interpretation capabilities are integrated in one system. When the system has the turn, its feedback prediction model could be used to identify places where feedback from the user could be expected. This information could be useful for differentiating nonverbal user behavior that is not relevant in terms of feedback (inconsequential head movements, self-touches) from behavior with very similar surface forms that is a feedback act. Furthermore, knowing when to expect feedback from a user could enable a more sophisticated interpretation of the absence of feedback (no feedback in certain places can be considered negative feedback). Conversely, when the user has the turn and the system is expected to provide feedback, it could use its own feedback interpretation system to predict the communicative effects of various feedback signals in the specific context in order to choose the right behavior.

6. CONCLUSIONS

In this paper, we have presented an overview of the literature in the field of feedback from agents to humans (Section 3) and from humans to agents (Section 4). For agent-to-user feedback, we conclude that a viable future direction is to move toward grounding-focussed rather than surface-focussed feedback, to improve the form and suitability of the actual signals, compared to the current models that prioritize timing. This is a direction that requires work both in the synthesis and multimodal realization of feedback signals and the modeling of the user's state. For user-to-agent feedback, we find that theoretical models of grounding exist, but require multimodal processing of the user's feedback, moving beyond simply text.

Finally, we conclude that user-to-agent feedback and agent-to-user feedback are both beneficial to systems. This is

because users expect a system that produces socially complex behaviors to understand socially complex feedback, and the underlying models of how feedback should be produced overlap with how feedback should be interpreted. Thus, future work should focus on scenarios where the agent and the user can both take the turn and speak or give feedback.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

AA and GS were supported by the Swedish Foundation for Strategic Research (SSF) project Co-Adaptive Human-Robot Interactive Systems (COIN). HB was supported by the German Research Foundation (DFG) in the Collaborative Research Center TRR 318/1 2021 Constructing Explainability (438445824). We also acknowledge the financial support of the German Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University for the article processing charge.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process*. 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Al Moubayed, S., Baklouti, M., Chetouani, M., Dutoit, T., Mahdhaoui, A., Martin, J.-C., et al. (2009). "Generating robot/agent backchannels during a storytelling experiment," in *Proceedings of the IEEE International Conference on Robotics and Automation* (Kobe), 3749–3754. doi: 10.1109/ROBOT.2009.5152572

Allen, J., and Core, M. G. (1997). *Draft of DAMSL: Dialog Act Markup in Several Layers*. Available online at: https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/

Allwood, J. (1988). "Om det svenska systemet för språklig återkoppling," in *Svenskans Beskrivning 16, Vol. 1*, eds P. Linell, V. Adelswärd, T. Nilsson, and P. A. Pettersson (Linköping: Linköping University; Tema Kommunikation), 89–106.

Allwood, J., and Cerrato, L. (2003). "A study of gestural feedback expressions," in *Proceedings of the 1st Nordic Symposium on Multimodal Communication* (Copenhagen), 7–22.

Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E., and Koppensteiner, M. (2007). The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behaviour simulation. *Lang. Resour. Eval.* 41, 255–272. doi: 10.1007/s10579-007-9056-2

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *J. Semant*. 9, 1–26. doi: 10.1093/jos/9.1.1

Axelsson, A., and Skantze, G. (2022). Multimodal user feedback during adaptive robot-human presentations. *Front. Comput. Sci.* 3:741148. doi: 10.3389/fcomp.2021.741148

Axelsson, N., and Skantze, G. (2019). "Modelling adaptive presentations in human-robot interaction using behaviour trees," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue* (Stockholm), 345–352. doi: 10.18653/v1/W19-5940

Axelsson, N., and Skantze, G. (2020). "Using knowledge graphs and behaviour trees for feedback-aware presentation agents," in *Proceedings of the 20th International Conference on Intelligent Virtual Agents* (Glasgow), 1–8. doi: 10.1145/3383652.3423884

Baur, T., Schiller, D., and André, E. (2016). "Modeling user's social attitude in a conversational system," in *Emotions and Personality in Personalized Services*, eds M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir (Basel: Springer), 181–199. doi: 10.1007/978-3-319-31413-6_10

Bavelas, J. B., Chovil, N., Coates, L., and Roe, L. (1995). Gestures specialized for dialogue. *Pers. Soc. Psychol. Bull.* 21, 394–405. doi: 10.1177/0146167295214010

Bavelas, J. B., Chovil, N., Lawrie, D. A., and Wade, A. (1992). Interactive gestures. *Discour. Process.* 15, 469–489. doi: 10.1080/01638539209544823

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *J. Pers. Soc. Psychol.* 79, 941–952. doi: 10.1037/0022-3514.79.6.941

Bavelas, J. B., Coates, L., and Johnson, T. (2002). Listener responses as a collaborative process: the role of gaze. *J. Commun.* 52, 566–580. doi: 10.1111/j.1460-2466.2002.tb02562.x

Benotti, L., and Blackburn, P. (2021). "Grounding as a collaborative process," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (Kyiv), 515–531. doi: 10.18653/v1/2021.eacl-main.41

Bevacqua, E. (2009). *Computational model of listener behavior for embodied conversational agents* (Ph.D. thesis). Université Paris, Paris, France.

Bevacqua, E. (2013). "Chapter 10: A survey of listener behaviour and listener models for embodied conversational agents," in *Coverbal Synchrony in Human-Machine Interaction*, eds M. Rojc and N. Campbell (Boca Raton, FL: CRC Press), 243–268. doi: 10.1201/b15477-11

Bevacqua, E., Heylen, D., Pelachaud, C., and Tellier, M. (2007). "Facial feedback signals for ECAs," in *Proceedings of the AISB'07 Annual Convention: Symposium on Language, Speech and Gesture for Expressive Characters* (Newcastle).

Bevacqua, E., Mancini, M., and Pelachaud, C. (2008). "A listening agent exhibiting variable behavior," in *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (Tokyo), 262–269. doi: 10.1007/978-3-540-85483-8_27

Bohus, D. (2007). *Error awareness and recovery in conversational spoken language interfaces* (Ph.D. thesis). Carnegie Mellon University, Pittsburgh, PA, United States.

Bohus, D., and Horvitz, E. (2009). "Models for multiparty engagement in open-world dialog," in SIGDIAL *'09: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (London), 225–234. doi: 10.3115/1708376.1708409

Bohus, D., and Rudnicky, A. I. (2005). "Constructing accurate beliefs in spoken dialog systems," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (San Juan), 272. doi: 10.1109/ASRU.2005.1566504

Boyd, A., Puri, R., Shoeybi, M., Patwary, M., and Catanzaro, B. (2020). "Large scale multi-actor generative dialog modeling," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Seattle, WA: Association for Computational Lingusitics), 66–84. doi: 10.18653/v1/2020.acl-main.8

Brennan, S. E., and Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: a theoretical framework. *Knowl. Based Syst.* 8, 143–151. doi: 10.1016/0950-7051(95)98376-H

Brunner, M.-L., and Diemer, S. (2021). Multimodal meaning making: the annotation of nonverbal elements in multimodal corpus transcription. *Res. Corpus Linguist.* 10, 63–88. doi: 10.32714/ricl.09.01.05

Buschmeier, H. (2018). *Attentive speaking. From listener feedback to interactive adaptation* (Ph.D. thesis). Faculty of Technology, Bielefeld University, Bielefeld, Germany.

Buschmeier, H., and Kopp, S. (2012). "Using a Bayesian model of the listener to unveil the dialogue information state," in *SemDial 2012: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue* (Paris), 12–20.

Buschmeier, H., and Kopp, S. (2014). "A dynamic minimal model of the listener for feedback-based dialogue coordination," in *Proceedings of the 18th*

Workshop on the Semantics and Pragmatics of Dialogue (SemDial) (Edinburgh), 17–25.

Buschmeier, H., and Kopp, S. (2018). "Communicative listener feedback in human–agent interaction: artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (Stockholm), 1213–1221.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., and Pelachaud, C. (2014). "Representing communicative functions in saiba with a unified function markup language," in *Proceedings of the 14th International Conference on Intelligent Virtual Agents* (Boston, MA), 81–94. doi: 10.1007/978-3-319-09767-1_11

Cassell, J. (2001). Embodied conversational agents: representation and intelligence in user interfaces. *AI Mag.* 22, 67–83. doi: 10.1609/aimag.v22i4.1593

Cassell, J., and Thórisson, K. R. (1999). The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Appl. Artif. Intell.* 13, 519–538. doi: 10.1080/088395199117360

Cathcart, N., Carletta, J., and Klein, E. (2003). "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (Budapest), 51–58. doi: 10.3115/1067807.1067816

Cerrato, L. (2005). "Linguistic function of head nods," in *Proceedings from the 2nd Nordic Conference on Multimodal Communication* (Gothenburg), 137–152.

Chiba, Y., Nose, T., and Ito, A. (2017). "Analysis of efficient multimodal features for estimating user's willingness to talk: comparison of human-machine and human-human dialog," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (Kuala Lumpur), 428–431. doi: 10.1109/APSIPA.2017.8282069

Chiba, Y., Nose, T., Ito, M., and Ito, A. (2016). Estimating the user's state before exchanging utterances using intermediate acoustic features for spoken dialog systems. *IAENG Int. J. Comput. Sci.* 43, 1–9.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620539

Clark, H. H., and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *J. Mem. Lang.* 50, 62–81. doi: 10.1016/j.jml.2003.08.004

Clark, H. H., and Schaefer, E. F. (1989). Contributing to discourse. *Cogn. Sci.* 13, 259–294. doi: 10.1207/s15516709cog1302_7

Comas, J., Aspandi, D., and Binefa, X. (2020). "End-to-end facial and physiological model for affective computing and applications," in *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (Buenos Aires), 93–100. doi: 10.1109/FG47880.2020.00001

Core, M. G., and Allen, J. F. (1997). "Coding dialogs with the DAMSL annotation scheme," in *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines* (Cambridge, MA).

Crook, P., and Marin, A. (2017). "Sequence to sequence modeling for user simulation in dialog systems," in *Proceedings of Interspeech 2017* (Stockholm), 1706–1710. doi: 10.21437/Interspeech.2017-161

de Kok, I. (2013). *Listening heads* (Ph.D. thesis). University of Twente, Enschede, Netherlands.

de Kok, I., and Heylen, D. (2011). "The MultiLis corpus-dealing with individual differences in nonverbal listening behavior," in *Proceedings of the 3rd COST 2102 International Training School* (Caserta), 362–375. doi: 10.1007/978-3-642-18184-9_32

de Kok, I., and Heylen, D. (2012). "A survey on evaluation metrics for backchannel prediction models," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog* (Stevenson, WA), 15–18.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). "SimSensei Kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris), 1061–1068.

Ding, Y., Pelachaud, C., and Artières, T. (2013). "Modeling multimodal behaviors from speech prosody," in *Proceedings of the 13th International Conference on Intelligent Virtual Agents* (Edinburgh), 217–228. doi: 10.1007/978-3-642-40415-3_19

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

Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Commun.* 50, 630–645. doi: 10.1016/j.specom.2008.04.002

Edlund, J., House, D., and Skantze, G. (2005). "The effects of prosodic features on the interpretation of clarification ellipses," in *Proceedings of Interspeech 2005* (Lisbon), 2389–2392. doi: 10.21437/Interspeech.2005-43

Ekman, P. (1993). Facial expression and emotion. *Am. Psychol.* 48, 384–392. doi: 10.1037/0003-066X.48.4.384

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., and Purver, M. (2015). "Feedback in conversation as incremental semantic update," in *Proceedings of the 11th International Conference on Computational Semantics* (London), 261–271.

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* 133, 694–724. doi: 10.1037/0033-2909.133.4.694

Fujimoto, D. T. (2007). Listener responses in interaction: a case for abandoning the term backchannel. *Bull. Osaka Jogakuin Coll.* 37, 35–54.

Galati, A., and Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Goodwin, C. (1986). Between and within: alternative sequential treatments of continuers and assessments. *Hum. Stud.* 9, 205–217. doi: 10.1007/BF00148127

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., et al. (2006). "Virtual rapport," in *Proceedings of the 6th International Conference on Intelligent Intelligent Virtual Agents* (Marina del Rey, CA), 14–27. doi: 10.1007/11821830_2

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). "Creating rapport with virtual agents," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris), 125–138. doi: 10.1007/978-3-540-74997-4_12

Gravano, A., Beňuš, Š, Chávez, H., Hirschberg, J., and Wilcox, L. (2007). "On the role of context and prosody in the interpretation of 'okay'," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (Prague), 800–807.

Gravano, A., and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* 25, 601–634. doi: 10.1016/j.csl.2010.10.003

Gravano, A., Hirschberg, J., and Beňuš, Š. (2012). Affirmative cue words in task-oriented dialogue. *Comput. Linguist.* 38, 1–39. doi: 10.1162/COLI_a_00083

Guntz, T., Balzarini, R., Vaufreydaz, D., and Crowley, J. L. (2017). "Multimodal observation and interpretation of subjects engaged in problem solving," in *Proceedings of the 1st Workshop on Behavior, Emotion and Representation: Building Blocks of Interaction.*

Gustafson, J., and Neiberg, D. (2010). "Prosodic cues to engagement in non-lexical response tokens in Swedish," in *Proceedings of the DiSS-LPSS Joint Workshop 2010* (Tokyo).

Hadar, U., Steiner, T. J., and Clifford Rose, F. (1985). Head movement during listening turns in conversation. *J. Nonverb. Behav.* 9, 214–228. doi: 10.1007/BF00986881

Hanna, N., and Richards, D. (2019). Speech act theory as an evaluation tool for human-agent communication. *Algorithms* 12:79. doi: 10.3390/a12040079

Hee, E., Artstein, R., Lei, S., Cepeda, C., and Traum, D. (2017). "Assessing differences in multimodal grounding with embodied and disembodied agents," in *5th European and 8th Nordic Symposium on Multimodal Communication* (Bielefeld).

Heldner, M., Edlund, J., and Hirschberg, J. (2010). "Pitch similarity in the vicinity of backchannels," in *Proceedings of Interspeech 2010* (Makuhari), 3054–3057. doi: 10.21437/Interspeech.2010-58

Heldner, M., Hjalmarsson, A., and Edlund, J. (2013). "Backchannel relevance spaces," in *Proceedings of Nordic Prosody XI* (Tartu), 137–146.

Heylen, D. (2006). Head gestures, gaze and the principle of conversational structure. *Int. J. Human. Robot.* 3, 241–267. doi: 10.1142/S0219843606000746

Heylen, D. (2008). *Modeling Communication With Robots and Virtual Humans*. Berlin: Springer 241–259.

Heylen, D., Bevacqua, E., Tellier, M., and Pelachaud, C. (2007). "Searching for prototypical facial feedback signals," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris), 147–153. doi: 10.1007/978-3-540-74997-4_14

Heylen, D., Kopp, S., Marsella, S. C., Pelachaud, C., and Vilhjálmsson, H. (2008). "The next step towards a function markup language," in *Proceedings of the 8th International Conference on Intelligent Virtual Agents* (Tokyo), 270–280. doi: 10.1007/978-3-540-85483-8_28

Hjalmarsson, A., and Oertel, C. (2012). "Gaze direction as a back-channel inviting cue in dialogue," in *Proceedings of the IVA 2012 Workshop on Realtime Conversational Virtual Agents* (Santa Cruz, CA).

Howes, C., and Eshghi, A. (2021). Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *J. Logic Lang. Inform.* 30, 331–362. doi: 10.1007/s10849-020-09328-1

Huang, L., and Gratch, J. (2012). "Crowdsourcing backchannel feedback: understanding the individual variability from the crowds," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog* (Stevenson, WA), 31–34.

Huang, L., Morency, L.-P., and Gratch, J. (2010a). "Learning backchannel predicition model from parasocial consensus sampling: a subjective evaluation," in *Proceedings of the 10th International Conference on Intelligent Virtual Agents* (Philadelphia, PA), 159–172. doi: 10.1007/978-3-642-15892-6_17

Huang, L., Morency, L.-P., and Gratch, J. (2010b). "Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems* (Toronto, ON), 1265–1272.

Hussain, N., Erzin, E., Sezgin, T. M., and Yemez, Y. (2019). "Speech driven backchannel generation using deep Q-network for enhancing engagement in human-robot interaction," in *Proceedings of Interspeech 2019* (Graz), 4445–4449. doi: 10.21437/Interspeech.2019-2521

Inden, B., Malisz, Z., Wagner, P., and Wachsmuth, I. (2013). "Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent," in *Proceedings of the 15th International Conference on Multimodal Interaction* (Sydney, NSW), 181–188. doi: 10.1145/2522848.2522890

Ishi, C. T., Ishiguro, H., and Hagita, N. (2014). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Commun.* 57, 233–243. doi: 10.1016/j.specom.2013.06.008

Johansson, M., Hori, T., Skantze, G., Höthker, A., and Gustafson, J. (2016). "Making turn-taking decisions for an active listening robot for memory training," in *Proceedings of the International Conference on Social Robotics* (Kansas City, MO), 940–949. doi: 10.1007/978-3-319-47437-3_92

Jokinen, K., Furukawa, H., Nishida, M., and Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst.* 3, 1–30. doi: 10.1145/2499474.2499481

Jokinen, K., and Majaranta, P. (2013). "Eye-gaze and facial expressions as feedback signals in educational interactions," in *Technologies for Inclusive Education: Beyond Traditional Integration Approaches*, eds D. Griol Barres, Z. Callejas Carrión, and R. López-Cózar Delgado (Hershey, PA: IGI Global), 38–58. doi: 10.4018/978-1-4666-2530-3.ch003

Jonsdottir, G. R., Gratch, J., Fast, E., and Thórisson, K. R. (2007). "Fluid semantic back-channel feedback in dialogue: challenges & progress," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris), 154–160. doi: 10.1007/978-3-540-74997-4_15

Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). "Lexical, prosodic, and syntactic cues for dialog acts," in *Proceedings of the ACL-COLING 1998 Workshop on Discourse Relations and Discourse Markers* (Montreal, QC), 114–120.

Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., and Ward, N. G. (2016). "Prediction and generation of backchannel form for attentive listening systems," in *Proceedings of Interspeech 2016* (San Francisco, CA), 2890–2894. doi: 10.21437/Interspeech.2016-118

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol.* 26, 22–63. doi: 10.1016/0001-6918(67)90005-4

Keysar, B. (1997). Unconfounding common ground. *Discourse Process.* 24, 253–270. doi: 10.1080/01638539709545015

Khosla, R., Chu, M.-T., Kachouie, R., Yamada, K., Yoshihiro, F., and Yamaguchi, T. (2012). "Interactive multimodal social robot for improving quality of care of elderly in Australian nursing homes," in *Proceedings of the 20th ACM International Conference on Multimedia* (Nara), 1173–1176. doi: 10.1145/2393347.2396411

Kleckova, J., Kral, P., and Krutisova, J. (2005). "Use of nonverbal communication in dialog system," in *Proceedings of the 4th WSEAS/IASME International Conference on System Science and Simulation in Engineering* (Teneriffe), 280–283.

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

Kontogiorgos, D., Pereira, A., and Gustafson, J. (2021). Grounding behaviours with conversational interfaces: effects of embodiment and failures. *J. Multim. User Interfaces* 15, 239–254. doi: 10.1007/s12193-021-00366-y

Kopp, S., Allwood, J., Grammar, K., Ahlsén, E., and Stocksmeier, T. (2008). "Modeling embodied feedback with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, eds I. Wachsmuth and G. Knoblich (Berlin: Springer-Verlag), 18–37. doi: 10.1007/978-3-540-79037-2_2

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., et al. (2006). "Towards a common framework for multimodal generation: the behavior markup language," in *Proceedings of the 6th International Conference on Intelligent Virtual Agents* (Marina del Rey, CA), 205–217. doi: 10.1007/11821830_17

Kopp, S., Stocksmeier, T., and Gibbon, D. (2007). "Incremental multimodal feedback for conversational agents," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris), 139–146. doi: 10.1007/978-3-540-74997-4_13

Kopp, S., and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Comput. Anim. Virt. Worlds* 15, 39–52. doi: 10.1002/cav.6

Krauss, R. M., and Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *J. Pers. Soc. Psychol.* 4, 343–346. doi: 10.1037/h0023705

Kulic, D., and Croft, E. A. (2007). Affective state estimation for human–robot Interaction. *IEEE Trans. Robot.* 23, 991–1000. doi: 10.1109/TRO.2007.904899

Laban, G., George, J. N., Morrison, V., and Cross, E. S. (2021). Tell me more! assessing interactions with social robots from speech. *Paladyn* 12, 136–159. doi: 10.1515/pjbr-2021-0011

Lai, C. (2010). "What do you mean, you're uncertain?: the interpretation of cue words and rising intonation in dialogue," in *Proceedings of Interspeech 2010* (Makuhari), 1413–1416. doi: 10.21437/Interspeech.2010-429

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., et al. (2019). An evaluation dataset for intent classification and out-of-scope prediction. *arXiv:1909.02027*. doi: 10.18653/v1/D19-1131

Larsson, S. (2003). "Interactive communication management in an issue-based dialogue system," in *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue* (Saarbrücken), 75–82.

Lee, J., and Marsella, S. C. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Trans. Multim.* 12, 552–562. doi: 10.1109/TMM.2010.2051874

Li, Y., Qian, K., Shi, W., and Yu, Z. (2020). "End-to-end trainable non-collaborative dialog system," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence, Vol. 34* (New York, NY), 8293–8302. doi: 10.1609/aaai.v34i05.6345

Lisetti, C. L., and Rumelhart, D. E. (1998). "Facial expression recognition using a neural network," in *Proceedings of the 11th International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (Sanibel Island, FL), 328–332.

Lisetti, C. L., and Schiano, D. J. (2000). Automatic facial expression interpretation: where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragm. Cogn.* 8, 185–235. doi: 10.1075/pc.8.1.09lis

Liu, B., and Lane, I. (2017). "Dialog context language modeling with recurrent neural networks," in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA), 5715–5719. doi: 10.1109/ICASSP.2017.7953251

Liu, Y., Han, K., Tan, Z., and Lei, Y. (2017). "Using context information for dialog act classification in DNN framework," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen), 2170–2178. doi: 10.18653/v1/D17-1231

Ma, Y., Zeng, Z., Zhu, D., Li, X., Yang, Y., Yao, X., et al. (2020). "An end-to-end dialogue state tracking system with machine reading comprehension and wide & deep classification," in *Proceedings of the AAAI-20 8th Dialog System Technology Challenge (DSTC8)* (New York, NY: AAAI). Available online at: https://arxiv.org/abs/1912.09297

Malisz, Z., Włodarczak, M., Buschmeier, H., Skubisz, J., Kopp, S., and Wagner, P. (2016). The ALICO corpus: analysing the active listener. *Lang. Resour. Eval.* 50, 411–442. doi: 10.1007/s10579-016-9355-6

Marechal, C., Mikołajewski, D., Tyburek, K., Prokopowicz, P., Bougueroua, L., Ancourt, C., et al. (2019). "Survey on AI-based multimodal methods for emotion detection," in *High-Performance Modelling and Simulation for Big Data Applications*, eds J. Kołodziej and H. González-Vélez (Cham: Springer), 307–324. doi: 10.1007/978-3-030-16272-6_11

McClave, E. Z. (2000). Linguistic functions of head movement in the context of speech. *J. Pragm.* 32, 855–878. doi: 10.1016/S0378-2166(99)00079-X

Meena, R., Skantze, G., and Gustafson, J. (2014). Data-driven models for timing feedback responses in a Map Task dialogue system. *Comput. Speech Lang.* 28, 903–922. doi: 10.1016/j.csl.2014.02.002

Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., and Nakamura, S. (2011a). "Analysis on effects of text-to-speech and avatar agent in evoking users' spontaneous listener's reactions," in *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop* (New York, NY: Springer), 77–89. doi: 10.1007/978-1-4614-1335-6_10

Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., and Nakamura, S. (2011b). "Toward construction of spoken dialogue system that evokes users' spontaneous backchannels," in *Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Portland, OR), 259–265.

Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Auton. Agents Multiagent Syst.* 20, 70–84. doi: 10.1007/s10458-009-9092-y

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2005). "Contextual recognition of head gestures," in *Proceedings of the 7th international conference on Multimodal Interfaces* (Trento), 18–24. doi: 10.1145/1088463.1088470

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2007). Head gestures for perceptual interfaces: the role of context in improving recognition. *Artif. Intell.* 171, 568–585. doi: 10.1016/j.artint.2007.04.003

Mueller, M., Leuschner, D., Briem, L., Schmidt, M., Kilgour, K., Stueker, S., et al. (2015). "Using neural networks for data-driven backchannel prediction: a survey on input features and training techniques," in *Proceedings of the 17th International Conference, HCI International* (Los Angeles, CA: Springer), 259–265. doi: 10.1007/978-3-319-20916-6_31

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., and Hagita, N. (2009). "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction* (La Jolla, CA), 61–68. doi: 10.1145/1514095.1514109

Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). "Towards a model of face-to-face grounding," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo), 553–561. doi: 10.3115/1075096.1075166

Norman, D. A. (1990). *The Design of Everyday Things*. New York, NY: Doubleday.

Novick, D. G., Hansen, B., and Ward, K. (1996). "Coordinating turn-taking with gaze," in *Proceeding of 3th International Conference on Spoken Language Processing* (Philadelphia, PA), 1888–1891. doi: 10.1109/ICSLP.1996.608001

Oertel, C., Lopes, J., Yu, Y., Funes Mora, K. A., Gustafson, J., Black, A. W., et al. (2016). "Towards building an attentive artificial listener: on the perception of attentiveness in audio-visual feedback tokens," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo), 21–28. doi: 10.1145/2993148.2993188

Ortega, D., and Vu, N. T. (2017). "Neural-based context representation learning for dialog act classification," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (Saarbrücken), 247–252. doi: 10.18653/v1/W17-5530

Ouyang, Y., Chen, M., Dai, X., Zhao, Y., Huang, S., and Chen, J. (2020). "Dialogue state tracking with explicit slot connection modeling," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Seattle, WA), 34–40. doi: 10.18653/v1/2020.acl-main.5

Paggio, P., Agirrezabal, M., Jongejan, B., and Navarretta, C. (2020). "Automatic detection and classification of head movements in face-to-face conversations," in *Proceedings of LREC2020 Workshop on People in Language, Vision and the Mind (ONION2020)* (Marseille), 15–21.

Paggio, P., Navarretta, C., and Jongejan, B. (2017). "Automatic identification of head movements in video-recorded conversations: can words help?" in *Proceedings of the 6th Workshop on Vision and Language* (Valencia), 40–42. doi: 10.18653/v1/W17-2006

Pammi, S. (2011). *Synthesis of listener vocalizations. Towards interactive speech synthesis* (Ph.D. thesis). Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany.

Petukhova, V., and Bunt, H. (2009). "Grounding by nodding," in *Proceedings of GESPIN-Gesture and Speech in Interaction* (Poznań).

Pichl, J., Marek, P., Konrád, J., Lorenc, P., Ta, V. D., and Sedivý, J. (2020). Alquist 3.0: Alexa prize bot using conversational knowledge graph. *CoRR, abs/2011.03261*.

Poggi, I., D'Errico, F., and Vincze, L. (2010). "Types of nods. The polysemy of a social signal," in *Proceedings of the 7th International Conference on Language Resources and Evaluation* (Valletta), 2570–2576.

Poppe, R., Truong, K. P., and Heylen, D. (2013). Perceptual evaluation of backchannel strategies for artificial listeners. *Auton. Agents Multiagent Syst.* 27, 235–253. doi: 10.1007/s10458-013-9219-z

Porhet, C., Ochs, M., Saubesty, J., De Montcheuil, G., and Bertrand, R. (2017). "Mining a multimodal corpus of doctor's training for virtual patient's feedbacks," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 473–478. doi: 10.1145/3136755.3136816

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. *Inform. Fus.* 37, 98–125. doi: 10.1016/j.inffus.2017.02.003

Prepin, K., Ochs, M., and Pelachaud, C. (2013). "Beyond backchannels: co-construction of dyadic stance by reciprocal reinforment of smiles between virtual agents," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (Berlin), 1163–1168.

Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., and Sheth, A. (2015). "Intent classification of short-text on social media," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)* (Chengdu), 222–228. doi: 10.1109/SmartCity.2015.75

Purver, M. (2004). *The theory and use of clarification requests in dialogue* (Ph.D. thesis). University of London, London, United Kingdom.

Qian, Y., Ubale, R., Ramanaryanan, V., Lange, P., Suendermann-Oeft, D., Evanini, K., and Tsuprun, E. (2017). "Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (Okinawa), 569–576. doi: 10.1109/ASRU.2017.8268987

Qin, L., Che, W., Li, Y., Ni, M., and Liu, T. (2020). DCR-net: a deep co-interactive relation network for joint dialog act recognition and sentiment classification. *Proc. AAAI Conf. Artif. Intell.* 34, 8665–8672. doi: 10.1609/aaai.v34i05.6391

Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., et al. (2011). Continuous interaction with a virtual human. *J. Multimodal User Interfaces* 4, 97–118. doi: 10.1007/s12193-011-0060-x

Rieser, V., and Lemon, O. (2011). *Reinforcement Learning for Adaptive Dialogue Systems. A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-24942-6

Rodríguez, K. J., and Schlangen, D. (2004). "Form, intonation and function of clarification requests in German task-oriented spoken dialogues," in *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue* (Barcelona), 101–108.

Ruede, R., Müller, M., Stüker, S., and Waibel, A. (2019). "Yeah, right, uh-huh: a deep learning backchannel predictor," in *Proceedings of the 8th International Workshop on Spoken Dialog Systems* (Siracusa), 247–258. doi: 10.1007/978-3-319-92108-2_25

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Lausanne), 305–311. doi: 10.1145/1957656.1957781

Schodde, T., Hoffmann, L., and Kopp, S. (2017). "How to manage affective state in child-robot tutoring interactions?" in *2017 International Conference on Companion Technology (ICCT)* (Ulm), 1–6. doi: 10.1109/COMPANION.2017.8287073

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., et al. (2012). Building autonomous sensitive artificial listeners.

IEEE Trans. Affect. Comput. 3, 165–183. doi: 10.1109/T-AFFC.2011.34

Schwarz, J., Marais, C. C., Leyvand, T., Hudson, S. E., and Mankoff, J. (2014). "Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, ON), 3443–3452. doi: 10.1145/2556288.2556989

Searle, J. R. (1969). *Speech Acts. An Essay in the Philosophy of Language*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9781139173438

Shi, H., Ushio, T., Endo, M., Yamagami, K., and Horii, N. (2017). "Convolutional neural networks for multi-topic dialog state tracking," in *Dialogues With Social Robots*, eds K. Jokinen and W. Graham (Singapore: Springer), 451–463. doi: 10.1007/978-981-10-2585-3_37

Shimojima, A., Koiso, H., Swerts, M., and Katagiri, Y. (1998). "An informational analysis of echoic responses in dialogue," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (Madison, WI), 951–956.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., et al. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. Speech* 41, 443–492. doi: 10.1177/002383099804100410

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., et al. (2018). A review of emotion recognition using physiological signals. *Sensors* 18:2074. doi: 10.3390/s18072074

Sidner, C. L., Lee, C., Morency, L.-P., and Forlines, C. (2006). "The effect of head-nod recognition in human-robot conversation," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, UT), 290–296. doi: 10.1145/1121241.1121291

Skantze, G. (2007). *Error handling in spoken dialogue systems. managing uncertainty, grounding and miscommunication* (Ph.D. thesis). Computer Science and Communication; Department of Speech, Music and Hearing; KTH Stockholm, Stockholm, Sweden.

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: a review. *Comput. Speech Lang.* 67, 101–178. doi: 10.1016/j.csl.2020.101178

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

Skantze, G., House, D., and Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialog. *Proc. Interspeech* 4, 2002–2005. doi: 10.21437/Interspeech.2006-548

Skantze, G., Johansson, M., and Beskow, J. (2015). "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, WA), 67–74. doi: 10.1145/2818346.2820749

Skantze, G., and Schlangen, D. (2009). "Incremental dialogue processing in a micro-domain," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (Athens), 745–753. doi: 10.3115/1609067.1609150

Stocksmeier, T., Kopp, S., and Gibbon, D. (2007). "Synthesis of prosodic attitudinal variants in German backchannel "ja"," in *Proceedings of Interspeech 2007* (Antwerp), 1290–1293. doi: 10.21437/Interspeech.2007-232

Thiebaux, M., Marshall, A., Marsella, S., and Kallmann, M. (2008). "SmartBody: behavior realization for embodied conversational agents," in *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems* (Estoril), 151–158.

Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi: 10.1207/s15327965pli0104_1

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Technical report, Rochester, NY: University of Rochester.

Traum, D. R., and Hinkelmann, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Comput. Intell.* 8, 575–599. doi: 10.21236/ADA256368

Truong, K. P., Poppe, R., de Kok, I., and Heylen, D. (2011). "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Proceedings of Interspeech 2011* (Florence), 2973–2976. doi: 10.21437/Interspeech.2011-744

Tzirakis, P., Chen, J., Zafeiriou, S., and Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Inform. Fus.* 68, 46–53. doi: 10.1016/j.inffus.2020.10.011

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Select. Top. Signal Process.* 11:1301-1309. doi: 10.1109/JSTSP.2017.2764438

van Welbergen, H., Reidsma, D., Ruttkay, Z. M., and Zwiers, J. (2009). Elckerlyc-A BML realizer for continuous, multimodal interaction with a virtual human. *J. Multimodal User Interfaces* 3, 271–284. doi: 10.1007/s12193-010-0051-3

Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., et al. (2007). "The behavior markup language: recent developments and challenges," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (Paris), 99–111. doi: 10.1007/978-3-540-74997-4_10

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: survey of an emerging domain. *Image Vis. Comput.* 27, 1743–1759. doi: 10.1016/j.imavis.2008.11.007

Visser, T., Traum, D. R., DeVault, D., and op den Akker, R. (2014). A model for incremental grounding in spoken dialogue systems. *J. Multimodal User Interfaces* 8, 61–73. doi: 10.1007/s12193-013-0147-7

Wallers, s., Edlund, J., and Skantze, G. (2006). "The effect of prosodic features on the interpretation of synthesised backchannels," in *Proceedings of the International Tutorial and Research Workshop on Perception and Interactive Technologies* (Kloster Irsee), 183–187. doi: 10.1007/11768029_19

Wang, Z., Lee, J., and Marsella, S. (2011). "Towards more comprehensive listening behavior: beyond the bobble head," in *Proceedings of the 11th International Conference on Intelligent Virtual Agents* (Reykjavík), 216–227. doi: 10.1007/978-3-642-23974-8_24

Wang, Z., Lee, J., and Marsella, S. (2013). Multi-party, multi-role comprehensive listening behaviour. *Auton. Agents Multiagent Syst.* 27, 218–234. doi: 10.1007/s10458-012-9215-8

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragm. Cogn.* 14, 129–182. doi: 10.1075/pc.14.1.08war

Ward, N. G. (1996). "Using prosodic clues to decide when to produce back-channel utterances," in *Proceedings of the 4th International Conference on Spoken Language Processing* (Philadelphia, PA), 1728–1731. doi: 10.1109/ICSLP.1996.607961

Ward, N. G., and DeVault, D. (2016). Challenges in building highly-interactive dialog systems. *AI Mag.* 37, 7–18. doi: 10.1609/aimag.v37i4.2687

Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal and the Machine, 2nd Edn.* Cambridge, MA: The MIT Press.

Williams, J. D., Henderson, M., Raux, A., Thomson, B., Black, A., and Ramachandran, D. (2014). The dialog state tracking challenge series. *AI Mag.* 35, 121–124. doi: 10.1609/aimag.v35i4.2558

Włodarczak, M., Buschmeier, H., Malisz, Z., Kopp, S., and Wagner, P. (2012). "Listener head gestures and verbal feedback expressions in a distraction task," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog* (Stevenson, WA), 93–96.

Xu, P., and Hu, Q. (2018). "An end-to-end approach for handling unknown slot values in dialogue state tracking," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Melbourne, NSW: Association for Computational linguistics), 1448–1457. doi: 10.18653/v1/P18-1134

Yankelovich, N., Levow, G.-A., and Marx, M. (1995). "Designing SpeechActs: issues in speech user interfaces," in *Proceedings of the 1995 SIGCHI Conference on Human Factors in Computing Systems (CHI)* (Denver, CO), 369–376. doi: 10.1145/223904.223952

Yngve, V. H. (1970). "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, eds M. A. Campbell, et al. (Chicago, IL: Chicago Linguistic Society), 567–577.

Zacharatos, H., Gatzoulis, C., and Chrysanthou, Y. L. (2014). Automatic emotion recognition based on body movement analysis: a survey. *IEEE Comput. Graph. Appl.* 34, 35–45. doi: 10.1109/MCG.2014.106

Zhang, J., Hashimoto, K., Wu, C.-S., Wang, Y., Yu, P., Socher, R., et al. (2020). "Find or classify? Dual strategy for slot-value predictions on multi-domain dialog state tracking," in *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics* (Barcelona), 154–167.

Zhou, J., Yu, K., Chen, F., Wang, Y., and Arshad, S. Z. (2018). "Multimodal behavioral and physiological signals as indicators of cognitive load," in *The Handbook of Multimodal-Multisensor Interfaces, Volume 2 Signal Processing, Architectures, and Detection of Emotion and Cognition*, eds S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Kruger (San Rafael, CA: Morgan & Claypool), 287–329. doi: 10.1145/3107990.3108002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.