frontiers in Computer Science

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fcomp.2021.741148

# Multimodal User Feedback During Adaptive Robot-Human Presentations

*Agnes Axelsson* and Gabriel Skantze**

*Department of Speech, Music and Hearing (TMH), KTH Royal Institute of Technology, Stockholm, Sweden*

Feedback is an essential part of all communication, and agents communicating with humans must be able to both give and receive feedback in order to ensure mutual understanding. In this paper, we analyse multimodal feedback given by humans towards a robot that is presenting a piece of art in a shared environment, similar to a museum setting. The data analysed contains both video and audio recordings of 28 participants, and the data has been richly annotated both in terms of multimodal cues (speech, gaze, head gestures, facial expressions, and body pose), as well as the polarity of any feedback (negative, positive, or neutral). We train statistical and machine learning models on the dataset, and find that random forest models and multinomial regression models perform well on predicting the polarity of the participants' reactions. An analysis of the different modalities shows that most information is found in the participants' speech and head gestures, while much less information is found in their facial expressions, body pose and gaze. An analysis of the timing of the feedback shows that most feedback is given when the robot makes pauses (and thereby invites feedback), but that the more exact timing of the feedback does not affect its meaning.

**Keywords: feedback, presentation, agent, robot, grounding, polarity, backchannel, multimodal**

***Edited by:***
*Hung-Hsuan Huang,*
*The University of Fukuchiyama, Japan*

***Reviewed by:***
*Raja Jamilah Raja Yusof,*
*University of Malaya, Malaysia*
*Paul Adam Bremner,*
*University of the West of England,*
*United Kingdom*

****Correspondence:***
*Agnes Axelsson*
*agnaxe@kth.se*
*Gabriel Skantze*
*skantze@kth.se*

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

***Received:*** *14 July 2021*
***Accepted:*** *06 December 2021*
***Published:*** *03 January 2022*

***Citation:***
*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*

## 1. INTRODUCTION

Agents communicating with humans must be able to both give and receive communicative feedback in order to ensure mutual understanding (Clark, 1996). While there has been a lot of work on how conversational agents should be able to provide feedback at appropriate places (Ward and Tsukahara, 2000; Poppe et al., 2010; Gravano and Hirschberg, 2011), less work has been done on how to pick up and interpret feedback coming from the user. To investigate this, we have in previous work explored the scenario of a robot presenting a piece of art in a shared environment, similar to a museum setting (Axelsson and Skantze, 2019, 2020). In such settings, the presenter can be expected to have the turn the majority of the time, while the listener (the audience) provides positive and negative feedback to the presenter. In our previous work, we have shown how the agent can use such feedback to adapt the presentation in real time, using behaviour trees and knowledge graphs to model the grounding status of the information presented (Axelsson and Skantze, 2020), and that an agent that adapts its presentation according to the feedback it receives is preferred by users (Axelsson and Skantze, 2019). However, since we have so far only evaluated the system using either a Wizard of Oz setup (Axelsson and Skantze, 2019) or with simulated users (Axelsson and Skantze, 2020), we have not yet addressed the question of how feedback in this setting could be automatically identified and classified. Identifying feedback is an important first step in creating an intelligent virtual agent for presentation. If a listener-aware system can identify and classify signals used by its audience as positive or negative, it opens up the possibility of using that classified feedback to create a highly adaptive, engaging agent.

A face-to-face setting, such as the one explored here, provides a wide range of different ways of expressing feedback across different modalities, including speech, head nods, facial expressions, gaze and body pose. Some of these modalities are harder to pick up and process by the agent than others, and some modalities might be complementary or redundant in terms of which information they carry. This calls for a more thorough analysis of how feedback is actually expressed in a robot-human presentation scenario, and what modalities are most important to process. In this article, we analyse a dataset of human-robot interactions recorded using a Wizard of Oz setup, to find out how humans spontaneously produce feedback toward a robot in these different modalities. Whereas much research on social signal processing is based on automatic analysis of the audio and video signals, we base this analysis on manually annotated features in the recorded data. This way, we can make sure that the findings are not dependent on the accuracy of any automatic detection of feedback signals.

Throughout this paper, we will look for answers to these questions:

1. What modalities are most commonly used to convey negative and positive feedback?
2. Are any modalities redundant or complimentary when it comes to expressing positive and negative feedback?
3. Does the interpretation of feedback as positive or negative change based on its relative timing to other feedback and the statement being reacted to?
4. Are there individual differences in the use of modalities to communicate different polarities of feedback?

This article is structured as follows. In section 2, we describe recent and basic work in the field of human-to-agent feedback, and feedback between humans in general. In section 3, we describe how we collected and annotated the data set used in this paper, and introduce the statistical models we use. Section 4 describes statistical patterns we found in the data as well as work on statistical models for approximating positivity and negativity based on multimodal signals. Section 5 is a discussion where we try to answer the questions listed above, and in section 6 we conclude and summarise the findings. The data we used for our study and analysis have been published; see section Data Availability Statement at the end of this paper.

# 2. RELATED WORK

## 2.1. Presentation Agents

Presentation agents (i.e., agents that present information to humans) can be seen as a sub-domain of conversational agents. However, whereas the initiative in general conversation can be mixed, the agent doing the presentation is expected to have the initiative most of the time, while paying attention to the listener's level of understanding or engagement. Kuno et al. (2007) found that mutual gaze and co-occurring nods were important indicators of an audience's engagement with a robot's presentation in a museum scenario. Recently, Velentza et al. (2020) found that a pair of presenting robots were more engaging than a single robot, but the embodiment of their robots—an Android tablet on top of a tripod—may make their results hard to apply to other, more embodied, scenarios. Iio et al. (2020) have shown that it is technologically feasible to have a robot walking around an enclosed exhibition at a museum. They present a robot that can identify individual visitors and use this identity information to adapt what the robot says, but the system is not interested in feedback from the users, beyond letting the users walk away if they are not interested in the presentation.

Another space where embodied presentation agents are used is the field of robot teachers. An argument in favour of robot teachers, proposed by Werfel (2014), is that they can be adaptive to the individual student while being as appealing as a human teacher to interact with. A typical teaching agent is the RoboThespian robot by Verner et al. (2016), which purely adapts to student responses in terms of which answer they choose on multi-choice questions; this is not an adaptive social agent, but rather a type of branching dialogue management. Tozadore and Romero (2020) presented a framework for how a virtual teaching agent can choose which questions to ask of a student depending on multimodal features—on a high level, the same types of right-or-wrong evaluations as Verner et al. (2016), but also more low-level features like facial features and attention estimated from gaze direction. Multimodal approaches for sensing student feedback in a school scenario do exist, even if they are not directly connected to a teaching robot; Goldberg et al. (2021) have presented proposals for multimodal machine-learning approaches for estimating individual students' engagement in classroom scenarios, but the approach may not extend beyond the specific setting.

A large body of work in connection to presentation agents relates to their ability to position themselves relatively to their audience in an actual museum scenario. While this is an important part of implementing an actual agent in the field, this track of research does not address the grounding of presented content toward the audience (Nourbakhsh et al., 2003; Kuzuoka et al., 2010; Yamaoka et al., 2010; Yousuf et al., 2012).

## 2.2. Feedback and Backchanneling in Communication

In the view of Yngve (1970), communication happens over a **main channel**, which carries the main message, as well as a **back channel**. Signals on the back channel—which have come to be called simply **backchannels** themselves—constitute feedback from the listener to the speaker. Feedback can be considered to be **positive** or **negative** (Rajan et al., 2001). This can be referred to as the **polarity** of the feedback (Allwood et al., 1992; Buschmeier and Kopp, 2018). On a wider scope, the polarity of entire utterances or statements is often referred to as *sentiment* (Wilson et al., 2009). Peters et al. (2005) gave the following example of multimodal feedback with a distinctly negative polarity:

> "For instance, to show you don't trust what is being said, a negative backchannel of believability, you can incline your head

> while staring obliquely and frowning to the Sender: two gaze signals combined with a head signal." (Peters et al., 2005)

A complementary view to that of Yngve (1970) was presented by Clark (1996), who instead split communication into *track 1* and *track 2*. The first track contains main contributions into the discourse, and the second track contains comments on content on the first track. Notably, this is not connected to turn-taking, unlike Yngve's main and back channel model: if the listener takes the turn and says "Wasn't his father dead by then?", then they have taken the turn, so the utterance is not part of the back channel, but the utterance is purely a comment on a previous utterance, so it is part of *track 2*.

Bavelas et al. (2000) proposed the difference between *specific* backchannels and *generic* backchannels. In this view, *specific backchannels* are direct comments on the context (e.g., frowning, "wow!") and *generic backchannels* are less specific signals whose main function is indicating that the listener is paying attention to the speaker's speech. Bavelas et al. (2002) also showed that gaze cues were an important signal from the speaker to invite backchannel activity from the listener. The view of backchannels as a tool by which the listener can shape the story of the speaker has been supported by a corpus study by Tolins and Fox Tree (2014). Their study showed that generic backchannels were likely to lead to the speaker continuing their story, while specific backchannels would lead to an elaboration or repair.

Clark (1996) described communication as a **joint project**, where both the speaker and listener give and receive feedback on several levels to fulfil the task of making the message come across. Clark and Krych (2004) showed that participants in an experiment were able to build a LEGO model more quickly when coordinating the building activity with an instructor, through speech and other modalities. This illustrates the parallel between a cooperative task and the communication used to facilitate the cooperative task. In the view of Clark (1994), communication is coordinated on four levels:

1. Vocalisation and attention
2. Presentation and identification
3. Meaning and understanding
4. Proposal and uptake

Each of these four levels forms a pair of actions on behalf of the speaker and the listener. Both actions must be performed at the same time for either one to be meaningful. The goal of an interaction is to reach mutual acceptance, where both participants believe that proposal and uptake have been achieved (Clark, 1996). Allwood et al. (1992) presented a similar model, where feedback serves primarily the four functions of indicating *contact*, *perception*, *understanding* and *attitudinal reactions*. In the model by Clark (1996), attitudinal reactions implicitly fall under *acceptance*, to the extent that they are covered by this model.

Feedback ladders like those presented by Clark (1996) create a system where polarised feedback on one level can imply feedback of the same or another polarity on another level. Clark (1996) defines two rules for how this process functions with the four levels mentioned above. *Upward completion* means that negative feedback on a low level of feedback implies negative feedback on all higher levels—negative *attention* implies negative *identification*, *understanding* and *uptake* since one can not identify what one is not paying attention to, can not understand what one has not identified, and can't accept what one has not understood (Clark, 1996). The inverse rule of *downward evidence* instead states that any feedback on a level—positively or negatively polarised—implies positive feedback on all lower levels. If one provides evidence of positive understanding, then this implies that the listener must also have identified and attended to the message (Clark, 1996). Notably, these two rules mean that *positive* feedback on a level says nothing about the levels *above* it—when the listener provides evidence of having positively understood an utterance, this neither says that they have accepted, nor that they have not accepted, that same utterance.

The **grounding criterion** is the level of feedback, among the four listed above, upon which feedback must be given for both the speaker and the listener to believe that communication works at any given point in time (Paek and Horvitz, 2000). For example, sometimes it might be enough that the listener shows continued attention, but sometimes the speaker might want to make sure that the listener has actually understood what has been said. If the speaker does not get enough feedback from the listener, the speaker might *elicit* feedback. The grounding criterion can change over the course of a conversation, or over the course of an individual utterance, as the speaker signals appropriate points in time for the listener to deploy backchannel signals, or elicits feedback on a higher level. Clark and Brennan (1991) also argued that the grounding criterion depends on the channels of communication being used for the discourse, with more limited methods of communication (phone, mail) requiring more explicit positive feedback than more multimodal methods of communication (face-to-face).

In the context of a presentation agent, the model of joint projects, joint problems and joint remedies presented by Clark (1996) can be a useful model for disambiguation between different types of feedback to the system, and a way to choose what strategies to use to repair problems in communication (Baker et al., 1999; Buschmeier and Kopp, 2013; Axelsson and Skantze, 2020). Buschmeier and Kopp (2011) argued that a Bayesian model, taking into account the previously estimated state of the user as well as the feedback as it is delivered, interpreted incrementally, is an appropriate method for a conversational agent to estimate the polarity and grounding level of the user's feedback at any given point in time. The advantage of such a model is that an absence of feedback can be represented as the user not having provided evidence of any polarity, which opens up for elicitation strategies.

If a presentation agent can identify and classify feedback given by the user, there are several ways in which the agent can use it to adapt the presentation. In Axelsson and Skantze (2020), we used a knowledge graph to keep track of the grounding status of specific facts presented to the user, creating a direct link between grounding and what statements are possible to present, as well as how the robot can refer to entities in the presentation. An alternative approach to this method is the approach presented by Pichl et al. (2020), where edges were inserted in a knowledge graph, containing information about the user's attitude and understanding of concepts. Alternative ways to adapt to identified and classified feedback have been presented by Buschmeier and Kopp (2011).

## 2.3. Feedback in Different Modalities

Feedback from the listener toward the speaker can be expressed in different modalities. Vocal feedback uses the auditory channel, and has both a verbal/linguistic component (the words being spoken), as well as non-verbal components, such as prosody (Stocksmeier et al., 2007; Malisz et al., 2012; Romero-Trillo, 2019). Non-vocal, non-verbal feedback is expressed in the visual channel (Jokinen, 2009; Nakatsukasa and Loewen, 2020), and can take the forms of gestures (Krauss et al., 1996), gaze (Kleinke, 1986; Thepsoonthorn et al., 2016), facial expressions (Buck, 1980; Krauss et al., 1996) and pose (Edinger and Patterson, 1983). In this section, we will provide a more thorough discussion on previous research related to feedback in those modalities.

### 2.3.1. Speech

A common form of vocal feedback are *backchannels*, like "uh-huh" (Yngve, 1970). There is also a span of vocal feedback that takes place somewhere between the main channel and the back channel. A specific form of such feedback is the **clarification request**, defined by Purver (2004) as a "dialogue device allowing a user to ask about some feature (e.g., the meaning or form) of an utterance, or part thereof." A similar notion is that of **echoic responses** (e.g., "tomorrow?"), where the listener repeats part of the speaker's utterance as a backchannel. These may serve as either an acknowledgement that the listener has heard that specific part of the speaker's utterance, or a repair request, where the original speaker must make an effort to clarify the previous utterance. Whether these should be considered negative or positive depends on whether they are interpreted as questions (i.e., a request for clarification) or not, and this difference can (to some extent) be signalled through prosody. The most commonly described tonal characteristic for questions is high final pitch and overall higher pitch (Hirst and Di Cristo, 1998), and this is especially true when the word order cannot signal the difference on its own. Several studies of fragmentary clarification requests (i.e., which signal negative feedback) have shown that they are associated with a rising final pitch, in both Swedish (Edlund et al., 2005) and German (Rodríguez and Schlangen, 2004).

The distinction between positive and negative feedback is also similar to the notion of **go on** and **go back** signals in dialogue, as proposed by Krahmer et al. (2002). In an analysis of a human-machine dialogue corpus, they found that *go back* signals were longer, lacked new information and often contained corrections or repetitions. They also found that there is a strong connection between the prosody and timing of the listener's response and whether the response is interpreted as *go on* or *go back*. Additionally, Krahmer et al. (2002) pointed out that the classification as *go on* and *go back* was dependent on the dialogue context; if the system says "Should I repeat that?" or "Did you understand?", the meaning of answers like "yes" or "no" can depend entirely on prosody and timing. Even when extended to more classes than *go on* and *go back*, like in the feedback classification schemes by Clark (1996) or Allwood et al. (1992) presented in section 2.2, the argument that context and multimodal signals (beyond pure linguistic content) can help distinguish between minimal pairs still holds.

Negative feedback can also be linked to the notion of "uncertainty," i.e., signs of uncertainty can also be regarded as negative feedback on some of the levels of understanding discussed in section 2.2. Skantze et al. (2014) explored user feedback in the context of a human-robot map task scenario, where the robot was instructing the users on how to draw a route. They showed that participants signalled uncertainty in their feedback through both prosody and word choice (lexical information). Uncertain utterances were shown to have a flatter pitch contour than certain utterances, and were also longer and had a lower intensity. Hough and Schlangen (2017) presented a grounding model for human-robot interaction where the robot could signal its uncertainty about what the user was referring to. The scenario was a pentamino block game, where the robot's goal was to identify the piece that the human referred to. Hough and Schlangen (2017) showed that users picked up the robot's uncertainty especially when it communicated it by moving more slowly toward the piece it thought the user was referring to. Hough and Schlangen (2017) framed uncertainty as a measure of the agent's certainty and understanding of its actions, as opposed to its knowledge—this is an extension of grounding.

### 2.3.2. Gaze

Gaze can be used for feedback from a listener toward a speaker, but this typically happens in combination with other signals and modalities. Mehlmann et al. (2014) showed that gaze is used by listeners to show that they understand which object the speaker is referring to, and that this co-occurs with the physiological process of actually finding the object. This also serves as a signal of *joint attention* from the listener toward the referred object. Gaze is also a way to improve the user's perception of the human-ness and social competence of the system (Zhang et al., 2017; Kontogiorgos et al., 2021; Laban et al., 2021).

Gaze is sometimes considered a gestural backchannel (Bertrand et al., 2007), although it is more commonly considered to be a turn-taking indication or turn-taking cue (Skantze, 2021). The uses of gaze as a turn-taking cue are not directly relevant to this paper, as our scenario has a strict turn-taking protocol, where the robot is the main speaker and the user mostly provides brief feedback.

On a low feedback level, mutual gaze can be used as a sign that a user wants to interact with the system, a signal called *engagement* by Bohus and Rudnicky (2006). Kuno et al. (2007) found that mutual gaze and co-occurring nods were important indicators of an audience's engagement with a robot's presentation in a museum scenario. Nakano and Ishii (2010) presented a model of gaze as a sign of mutual engagement. An agent that used a more sophisticated gaze sensing model was found to be preferrable to test participants.

### 2.3.3. Head Movements

Similarly to the results related to gaze by Bavelas et al. (2002) and the results related to prosody by Ward and Tsukahara (2000), McClave (2000) has shown that head-nods are a viable listener response to a backchannel-inviting cue from the speaker, especially if that cue is also a nod. Stivers (2008) presented the view that nods are a stronger signal than conventional backchannels like "uh-huh" and "OK," arguing that they present evidence that the recipient (listener) is able to visualise being part of the event being told by the teller (speaker).

Heylen (2005) presented a list of head movements together with the communicative functions they often serve, both for speakers and listeners. Heylen (2005) argued that head movements can be a communicative signal on both *track 1* and *track 2* as defined by Clark (1996) (see section 2.2). However, the examples presented by Heylen all related to head movements produced by the speaker, and were co-ordinated with speech or utterances that are unambiguously part of *track 1*— such as when the speaker produces an *inclusive* sweeping hand movement while saying the word "everything," indicating that the scope of the word is wide. Other gestures and functions listed by Heylen, like nodding to signal agreement, were more unambiguously part of *track 2*.

In a multimodal study of the *ALICO* corpus, Malisz et al. (2016) found that listeners used head movements twice as often as speech in response to being told a story by a speaker. Additionally, nods are by far the most common head movement feature, and multiple nods are twice as common as single nods. Similarly, head shakes occur much more often in multiples than one-by-one, but the opposite holds for head tilts and head jerks, which are significantly more likely to occur one-by-one than in multiples. Singh et al. (2018) arrived at different rates of usage of modalities when annotating a corpus of children reacting to each other's stories, finding that gazing on the speaker, smiling, leaning toward the speaker, raising one's brow, and responding verbally are all more common behaviours than nodding. While their results may not extend to adults, Singh et al. (2018) also found that adult evaluators considered certain signals to be indicative of positive or negative attention based on context—nodding was correlated with positive attention if the nod was long, but with negative attention if the nod was fast. Oertel et al. (2016) showed that head-nods are perceived as less indicative of attention the less pronounced they are, the slower they are, and the shorter they are, and conclude that head-nods are not merely a signal of positive attention, but rather reflect various degrees of attentiveness. The apparent difference between these results and those of Singh et al. (2018) could be argued to be because Singh et al. studied children, whose gestural behaviour is known to be different from that of adults (Colletta et al., 2010).

Navarretta et al. (2012) found that multiple nods are more common than single nods in Swedish and Danish experiment participants, while Finnish participants used single nods more often than multiple nods. This highlights how signals can be used differently even in cultures that are closely related to each other.

Novick (2012) showed that head-nods in dyadic conversations between humans were significantly cued by gaze—i.e., the listener would nod when gazed at by the speaker. This allows the listener to use feedback when it is the most likely to be picked up by the speaker, in a modality fitting for this. Novick and Gris (2013) showed that nods were less frequent in multi-party conversations, where there was more than one listener, and not necessarily cued by gaze.

Sidner et al. (2006) showed that there was a significant effect on how many head nods users of a system used *after* they figured out that the system could recognise such signals. These results generally go together with the findings by Kontogiorgos et al. (2021) and Laban et al. (2021), showing that this applies for head movements and speech, respectively.

### 2.3.4. Facial Expressions

Facial expressions are typically viewed as a sign of the listener's emotional state (Mehlmann et al., 2016): for example, eyebrow movements are known to signal interest or disbelief from a listener toward a speaker (Ekman, 2004). This is grounding on a high level—as mentioned in section 2.2, Allwood et al. (1992) placed attitudinal reactions as the highest level of the feedback scale, while Clark (1996) would classify it as a variant of showing acceptance.

Jokinen and Majaranta (2013) argued that facial expressions are closely tied to gaze signals; when interacting with a human, or with an embodied agent, listeners tend to gaze at the speaker's eyes and upper face region to be prepared to catch subtle facial expressions.

### 2.3.5. Body Pose

Body pose is typically only used as a unimodal indication of whether the sensed individual wants to engage with the system or not, as explored by Bohus and Rudnicky (2006). This is also how body pose was used in the model for estimating classroom engagement presented by Goldberg et al. (2021). Engagement corresponds to the lowest level of Clark's four levels of feedback described in section 2.2: *attention*. An exception to this is *shrugging*, which Goldberg et al. (2021) found to be used by children toward a reading partner agent. Shrugging was found by Goldberg et al. (2021) to signal that the child does not know the answer to a question. We would argue that this implies positive *identification* but negative or ambiguous *understanding* by the scheme detailed in section 2.2. Battersby (2011) showed that speakers were significantly more likely to use hand gestures than listeners.

Oppenheim et al. (2021) recently showed that leaning was used as an extended gaze cue by participants in an experiment where they had to learn how to build an object from another participant. In this experiment, the learners, corresponding to our listeners, coordinated gaze cues by leaning around 40% of the time when being taught how to build a smaller Lego model, and 26% of the time when being taught how to build a larger pipe structure—while the responses appeared to correspond to the physical properties of the object, indicating that learners' responses were cued by physically wanting to see the object being referred to, the gaze and lean signals happened in response to inviting cues by the teacher, indicating that the signals both served a grounding purpose and a practical purpose, simultaneously.

Park et al. (2019) found that there was a strong connection between the body pose of children— specifically the gesture of leaning forwards—and their intent to engage with a system.

Zaletelj and Košir (2017) have presented models for predicting classroom engagement from body pose in this sense. Body pose can also be used to predict the emotional state of a user: Sun et al. (2019) used body pose as an input to estimate subjects' emotional state, training a neural network on a dataset labelled with the speaker's intended emotional state and the listener's interpretation. This was then used in a robot that was consistently evaluated as more emotionally aware than a baseline.

#### 2.3.6. Combining Modalities

It important to not just study feedback functions of individual modalities, but to also consider their combined effects. Clark and Krych (2004) showed how multimodal grounding worked in an instructor-instructee scenario where a LEGO model was constructed by one participant. The data showed many multimodal patterns in how participants coordinated their speech with other modalities to ensure grounding (often specifically establishing which LEGO piece the speaker was referring to). Crucially, participants used visual modalities (for example, holding up a LEGO piece) when that modality resulted in an easier and faster reference than speech.

In a corpus study focusing on physiological indications of attention, Goswami et al. (2020) found that the rate of blinking, pupil dilation, head movement speed and acceleration, as well as prosodic features and facial landmarks can create a good model for predicting children's engagement with a task as well as when they are going to deploy backchannels. The authors' random forest model prioritises gaze as the most important feature for measuring whether the children were engaged with the task.

Visser et al. (2014) presented a model for how a conversational system could show grounding to a speaking user. This is the inverse of the scenario analysed by us, but the approach, where specific backchannels were assigned to specific states of hierarchic subcomponents of the system, is interesting. For example, their agent nodded if the language understanding component of the system reported a high confidence, and frowned if there was a pause longer than 200 ms and the language component was not confident that it had understood the last thing the speaker said. Kontogiorgos et al. (2019) presented a model of estimating uncertainty of a test participant by combining gaze and pointing modalities, but the authors conclude that it is uncertain how the results extend to domains outside of the specific test scenario.

Oppenheim et al. (2021) showed that the modalities that a listener used depended on the context of the cue used by the speaker. The scenario was a teacher/student scenario where the participants took turns teaching each other how to build an object. Depending on if the teacher looked at the student to *supplement*, *highlight* or *converse*, the student's response modalities significantly changed. Nod responses were significantly more common than speech responses if the teacher's act was *supplement*, nods and speech were approximately as common when responding to *highlight* actions, and speech was significantly more common in response to *converse* acts.

Hsieh et al. (2019) used multimodal features, specifically speech and head movements, to estimate the certainty of users of their virtual agent. They used *certainty* and *uncertainty* as a term for feedback that can be mapped to several of the levels we described in section 2.2. While the authors showed that applying statistical models to the data is a viable way to estimate self-reported certainty in the answers the users gave to the robot's questions, the study was limited by the small number of participants.

To conclude this review, a large body of work exists that investigates how individual modalities can be sensed and interpreted in terms of feedback. There is less work that investigate the combined effect of several modalities as feedback to achieve some task-specific goal between an agent and a user. More specifically, we have not found any previous systematic analyses of how human listeners express feedback in various modalities, as a response to a presentation agent.

## 3. METHOD

### 3.1. Data Collection

For our analysis, we set up an experiment where participants interacted with a robot presenting a piece of art, as seen in **Figure 1**, similar to that used in Axelsson and Skantze (2019). As a robot platform, the Furhat robot head was used, which has a back-projected animated face and a mechanical neck (Al Moubayed et al., 2012). The robot presented two paintings[1] to each participant. Although the agent communicated with the participants in English, participants were allowed to respond in Swedish or English if they wished. A Wizard of Oz setup was used, and the Wizard sat behind a separating wall on the other side of the room. Participants were led to believe that they were interacting with a fully autonomous system, and that the Wizard's role was only to make sure that data were being successfully recorded. The presentation was automated to a large extent, but the Wizard controlled whether the system would repeat utterances, move on, or clarify. To trigger more negative feedback (for the sake of the analysis), the system would misspeak with a certain probability (also controlled by the Wizard). Misspeaking was implemented by replacing a key part of the utterance with something else or by muting parts of the synthesised speech.

In total, 33 test participants were recruited. For technical reasons, data from 5 participants had to be discarded, leaving 28 usable participants. The participants were recorded on video from multiple angles, and were equipped with a microphone which recorded their speech. A snapshot from the video recording can be seen in **Figure 1**.

### 3.2. Manual Data Annotation

Video and audio data recorded from the experiment were synchronised and annotated in multiple ways. We cut the data into **clips** representing the time between the robot beginning an utterance and the video frame before it started its next utterance; this left us with between 59 and 125 clips per painting presentation. For the rest of this paper, **clip** will be used to refer to a video recording of the robot saying something, and the full

[1] The first painting presented to each participant was Pieter Brueghel's *Tower of Babel*, and the second was Gentile Bellini's *Miracle of the Cross fallen into the channel of Saint Lawrence*.

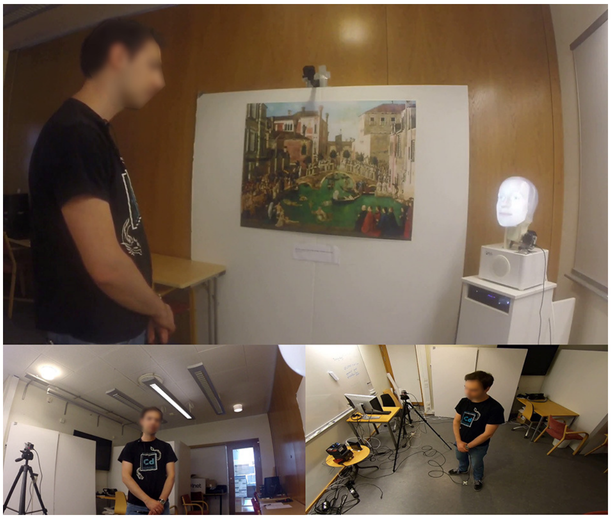

**FIGURE 1 |** A snapshot of an experiment participant, shown from all three angles from which we recorded them. The participant's face has been blurred for privacy. Top: the participant is shown from behind, illustrating the set-up of our experiment, with the presenting Furhat head visible.

reaction of the test participant to that utterance before the robot starts saying its next line.

### 3.2.1. Feedback Polarity of Each Clip

Each clip where the participant's turn was at least five seconds long was annotated based on what type of feedback it contained, using Amazon Mechanical Turk. Our original idea was to classify the feedback as being positive or negative acceptance, understanding, hearing or attention, as defined by Clark (1996) (discussed in section 2.2 above). However, initial results from annotating our data using these labels gave agreement scores that we considered too low to annotate the entire dataset this way. We believe that the scientific definitions of acceptance, understanding, hearing and attention were too nuanced and theoretical for laypeople to apply consistently, and that our annotators were mostly annotating clips as positive or negative attention, hearing, understanding or acceptance based on their immediate understanding of those words. This would have made the annotations questionable even if we were able to get a higher agreement score.

As a result of this, we decided to instead use the labels **positive**, **negative**, and **neutral**, and ignore the grounding level with which the feedback could be associated. The *positive* and *negative* labels were described to the annotators equivalently to the *go on* and *go back* labels used by Krahmer et al. (2002), with the *neutral* label representing cases where the annotator did not think that the participant's response was strong enough to classify it as either of the others:

- "Positive" was described as "Pick this if you think the person is showing that the presentation can continue the way it is currently going, without having to stop to repair something that went wrong. If the person shows understanding of what the robot said, or asks a followup question, then this may be the right option."
- "Negative" was described as "This option is right if you think the person is showing that the robot needs to stop and repeat something, or that it needs to explain something that the person didn't understand, hear or see."
- "Neutral" was described as "If the person doesn't really react or show any clear signals, or hasn't really had time to react by the end of the video, you should pick this."

**TABLE 1 |** The distribution of output features across all clips.

| Positivity | Negativity | Neutrality | Count | Argmax classification |
|---|---|---|---|---|
| 0 | 0 | 3 | 190 | Neutral |
| 0 | 1 | 2 | 109 | Neutral |
| 0 | 2 | 1 | 51 | Negative |
| 0 | 3 | 0 | 202 | Negative |
| 1 | 0 | 2 | 146 | Neutral |
| 1 | 1 | 1 | 74 | Neutral |
| 1 | 2 | 0 | 74 | Negative |
| 2 | 0 | 1 | 123 | Positive |
| 2 | 1 | 0 | 120 | Positive |
| 3 | 0 | 0 | 1,034 | Positive |

Each clip was annotated by three crowdworkers. In effect, this classified each clip by the *polarity* of the feedback contained in it, as defined in section 2.2.

**Table 1** shows how often each combination of positive, negative, and neutral evaluations appear. If the three evaluations are viewed as votes for a class, then the most common class is three votes for positivity (49%), followed by three votes for negativity (10%) and three votes for neutrality (9%). The final label for each clip in our dataset is determined based on the majority vote. 74 clips receive one vote for each class, and these are assigned the neutral label.

The distributions seen in **Table 1** give a Fleiss' $\kappa$ value of $\kappa \approx 0.582$, which is *moderate agreement* on the scale by Landis and Koch (1977). For a classification or annotation task, higher $\kappa$ values than this could be beneficial, but we accept this value since there are likely grey zones between the classes, and it is not obvious that clips with very few signals belong to any of the three classes without knowing the context of the clip. One annotator may assume that no signals are a positive signal, presuming that the context before the clip started is such that the listener has established a low grounding criterion for continuing the dialogue. Another annotator may have seen other clips of the robot eliciting feedback from users, and assume that the robot always wants the listener to react in some way, and thus evaluates the same clip as negative. For comparison, Malisz et al. (2016), who classified dialogues using four feedback levels similar to the schemes described in section 2.2, achieved a $\kappa$ score of around 0.3.

### 3.2.2. Multimodal Signals

Separately from the Mechanical Turk polarity classification, we also annotated each clip with what multimodal signals were used by the participant over time. This was done by a small number of annotators employed at KTH. We based our annotation system on the MUMIN standard by Allwood et al. (2007). To stay consistent with the clip format that had been used on Mechanical Turk, the multimodal feature annotation was also done clip-by-clip, rather than for the entire recording of a participant's interaction with the robot. This had the downside of making signals that were cut off by the beginning or end of a clip impossible to annotate properly, and the advantage of ensuring that only signals that could be fully seen by the Mechanical Turk evaluators were annotated.

In the MUMIN standard (Allwood et al., 2007), multimodal expressions are only annotated if they have a communicative function, and ignored if they are incidental to the communication (for example, if a participant blinks because their eyes are dry). MUMIN standardises how signals should be annotated for their feedback function (contact and perception, and optionally understanding), but we did not use this part of the standard, since we assumed that such information was captured in the polarity annotation by the Mechanical Turk crowdworkers. Finally, MUMIN is intended to be used for annotation of face-to-face interactions between two participants where the signals used by both participants are relevant and where turn-taking is important. However, since our scenario was very restricted in terms of turn-taking, and since we knew which behaviours were used by the speaking robot, we restricted our annotation to only use MUMIN constructs for annotating the signals used by the listening user, ignoring their meaning.

## 3.3. Data Processing

We post-processed and merged the annotated data from Mechanical Turk and our internal MUMIN-like annotation. Since the goal of this analysis is to find out what feedback could be detected in an online interaction, the post-processing involved taking steps to make the multimodal data more feasible to have been generated in real time. For example, two of the features annotated by our multimodal annotation were *multiple nods* and *single nod*—but it would be impossible for a real-time system to know if a nod is the first of many or a single nod in isolation at the time that the head gesture starts. To address this for head gestures and speech, our post-processing procedure involved replacing all head gesture features by an arbitrary *ongoing head gesture* feature. The feature that describes what the head gesture had been is only delivered on the last frame of *ongoing head gesture*. The same was done to speech (*ongoing speech*).

As a separate part of the post-processing process, we converted the transcribed text of the participants' speech to binary features representing the contents of the speech: *can't hear*, *can't see*, *no* and *yes*. These features were based on whether the transcribed text contains some variation on those phrases in either Swedish or English. The prosody of the speech segment was also converted into *rising F0* or *falling F0* through Praat (Boersma and van Heuven, 2001) by comparing the average *F0* for the first half of the speech segment to the *F0* in the second half, in the cases where Praat was able to extract these values. As mentioned in the previous paragraph, these classifications of speech contents or prosody are only delivered on the last frame of the *ongoing speech* annotation.

Our annotators also labelled speech as backchannels or non-backchannels. This distinction is not universal, and our labelling corresponds to what Duncan and Fiske (1977) would instead call *short back-channels* and *long back-channels*. Additionally, each speech segment was transcribed (if intelligible). Our annotators disagreed on whether short speech like *yeah* or *OK* should be considered backchannels or not, with a tendency to annotate them as non-backchannels. We retroactively went through the data and changed any speech segment that had been transcribed as just the single word *yes*, *yeah*, *OK*, *okay*, *yep*, and the corresponding phrases in Swedish, to backchannels if they had been annotated as non-backchannels. Longer speech segments containing the words in question (e.g., "OK, that makes sense.") were left as non-backchannels.

## 3.4. Statistical Models

Apart from analysing how the multimodal signals correlate with the three feedback labels (positive, negative, and neutral), we also apply four statistical models to our dataset, in order to analyse to what extent it is possible to predict these labels from the signals. We here provide a brief overview of these models.

### 3.4.1. Random Forest

Random forest models are a variant of decision tree models where a number of trees classify the data. If used for classification, like in our case, the majority vote determines the forest's classification. Random forests were originally proposed by Svetnik et al. (2003). They have previously performed well on feedback analysis tasks, like in the recent work by Jain et al. (2021), who recently successfully used random forests to identify multimodal feedback in clips of test participants, or in the work by Soldner et al. (2019), who successfully used random forests to classify whether participants in a study were lying based on multimodal cues. Yu and Tapus (2019) used random forests to classify emotions based on the combined modalities of thermal vision and body pose, finding that the random forest model successfully combined the modalities to achieve better performance than on either modality in isolation.

### 3.4.2. RPART Tree

RPART, short for *recursive partitioning*, is an algorithm for how to split the data when generating a decision tree. The trees are thus simply decision trees, and RPART is a name for the algorithm used to generate them (Hothorn et al., 2006).

RPART trees were not included in our analysis because we expected them to outperform random forest methods, but rather because they are easy to visualise and generate human-understandable patterns for classification. A brief analysis of the generated RPART trees can be found in section 4.2.3.

### 3.4.3. Multinominal Regression

Multinomial regression is an extension of logistic regression that allows for multi-class classification by linking the input signals to probabilities for each class. Multinomial regression and variants of logistic regression have been used successfully for dialogue state tracking (Bohus and Rudnicky, 2006) and multimodal signal sensing (Jimenez-Molina et al., 2018; Hsieh et al., 2019).

#### 3.4.4. LSTM Model

The LSTM neural network model, short for *long short-term memory*, was proposed by Hochreiter and Schmidhuber (1997). Neural networks utilising LSTMs have been used to model a large space of tasks since their introduction, including dialog state tracking (Zilka and Jurcicek, 2016; Pichl et al., 2020) and turn-taking (Skantze, 2017). Within the multimodal feedback space, Agarwal et al. (2019) have shown that LSTM models can perform incremental sentiment analysis, and Ma et al. (2019) have proposed model structures that make use of multimodal signals to classify emotions in subjects.

Our LSTM model starts with an embedding layer between the input features and the LSTM layer. The LSTM layer has 64 nodes, feeding into a three-wide embedding layer, feeding into a Softmax layer which gives the outputs as classification probabilities. Categorical cross-entropy is used as the loss function, and categorical accuracy as the accuracy function. For each fold, the model was trained for 100 epochs. The accuracy on the final time-step of each clip was calculated, and this accuracy value was used to choose the most accurate epoch. Larger and deeper models were tried out, but did not achieve better accuracy overall.

#### 3.4.5. Data Formatting

For the non-timing-aware random forest, RPART tree, and multinomial regression models, the multimodal feature set of each clip is converted into static features in four different ways. If the features are *split*, then the signals used during the robot's speech are separated from the features used during the participant's response. The alternative to this is *non-split*, where each feature represents the usage of a signal for the entire clip. If the features are formatted using *binary* formatting, then the value of the feature is 1 if is present at any point in the clip, and 0 otherwise; the alternative to this is *fractional* formatting, where a value between 0 and 1 (inclusive) is used, representing for how much of the clip the feature is present. This post-processing is required because only the LSTM model can be fed data by time-frame.

For the LSTM model, data is instead segmented into 100 ms time-frames, which allows it to make decisions that depend on the timing and ordering of the signals. In this data, a feature has the binary value of 1 if it is present at some point in the time-frame, and 0 otherwise, with the exceptions for specific head gestures and speech classification mentioned in section 3.3—speech was turned into *ongoing speech* until its final time-frame, and head gestures were turned into *ongoing head gesture*.

## 4. RESULTS

We split our results into two parts. In section 4.1, the data that we collected and annotated is analysed for statistical patterns. In section 4.2, we investigate to what extent it is possible to predict a clip's polarity from its multimodal signals using various statistical models.

### 4.1. Data Analysis

A summary of the signals that were annotated can be seen in **Table 2**. This table groups annotated signals by modality as **pose**, **facial expressions**, **gaze direction**, **head gestures**, and **speech**. The duration of the signal is not taken into account here. The *positive* class has a clear correlation with head nods, and the *negative* class has a clear correlation with the *speech was no* class, but the *neutral* class is mostly characterised by a lack of signals. It is never the class with the highest proportion of a signal, and when a signal appears more often in neutral clips than in positive *or* neutral clips, it is usually a signal that we would assume to be ambiguous, like *arms misc* or *head gesture was single tilt*.

#### 4.1.1. Individual Features Correlated With Labels

The rightmost two columns of **Table 2** show analyses of the distribution of feedback signals across clips labelled as positive, negative, and neutral. We use a $\chi^2$ test, Bonferroni corrected, to find if any signal is significantly more or less common in any of the three labels. If this is true, illustrated by the "Sign. 3" column in **Table 2**, we perform a follow-up test only on the positive and negative classes, and report $\chi^2$ test significance on that test as well. The results of the follow-up test are presented in the column labelled "Sign. ±."

Many signals, notably all speech signals and most head movement and facial gesture signals, show strong significance when comparing the distribution of all three labels to the distribution of the signal. Looking at the percentages of how often the signal shows up across different labels, discussed above, we find that strong significance in "Sign. 3" typically means that the signal is a strong indication that the clip is not neutral. This is clearly the case for speech and its sub-signals, where speech appears more than 70% of the time in positive or negative clips, but only 7.5% of the time in neutral clips.

An exception to significance in the "Sign. 3" column indicating that the clip is not neutral appears to be the "Jerk forwards" head gesture feature. This feature appears more often in neutral clips than in positive clips. Because of this, the significance in the "Sign. 3" column can be seen as evidence that this signal shows only that a clip is not positive, but that it could still be neutral.

Some signals also have strong significance when comparing the distribution between only positive and negative clips, presented in the rightmost column labelled "Sign ±." Since this is only a comparison of two classes, a quick *post-hoc* test can be performed by simply comparing the percentages of how often the signals appear in positive and negative videos. This is indicated in **Table 2** by marking the most common label in bold, where "Sign ±" is significant.

Frowning is significantly connected to negativity, since it appears much more often in negative clips than in positive clips. Speech in general only indicates that a clip is not neutral, but the sub-classifications of speech are strongly correlated with positivity or negativity. The "no" and "yes" features are strongly correlated with negativity and positivity, respectively. Rising F0 is connected to negativity, which can be explained by its

**TABLE 2 |** How common each signal detailed in section 3.3 is, clip-by-clip, for all clips, positive clips, negative clips and neutral clips.

| Modality | Signal | All (%) | Positive (%) | Negative (%) | Neutral (%) | Sign. 3$^{\dagger}$ | Sign. ±$^{\ddagger}$ |
|---|---|---|---|---|---|---|---|
| Pose | Cross arms | 9.2 | 9.1 | 11.9 | 7.9 | ns | ns |
| | Arms behind the back | 0.0 | 0.1 | 0.0 | 0.0 | ns | ns |
| | Arms misc | 31.8 | 33.1 | 28.2 | 31.1 | ns | ns |
| | Shrug | 0.4 | 0.4 | 1.0 | 0.0 | ns | ns |
| Face | Eyebrow raise | 9.8 | 11.6 | 12.2 | 4.3 | ** | ns |
| | Frown | 12.9 | 8.5 | **32.4** | 11.2 | **** | **** |
| | Facial laughter | 8.3 | 8.8 | 10.9 | 5.5 | ns | ns |
| | Lip pout | 4.8 | 6.0 | 5.1 | 1.8 | * | ns |
| | Mouth miscellaneous | 18.4 | 20.6 | 24.4 | 9.4 | **** | ns |
| | Smile | 25.5 | 29.7 | 34.0 | 10.4 | **** | ns |
| Gaze | Gaze on miscellaneous | 3.0 | 3.4 | 2.2 | 2.6 | ns | ns |
| | Gaze on poster | 98.6 | 98.8 | 98.1 | 98.2 | ns | ns |
| | Gaze on robot | 75.0 | 74.7 | 84.9 | 69.7 | ns | ns |
| Head gestures | Head gesture | 68.1 | **88.6** | 49.7 | 31.5 | **** | **** |
| | Jerk backwards | 2.6 | 2.7 | 4.2 | 1.6 | ns | ns |
| | Jerk forwards | 4.5 | 3.4 | **9.0** | 4.5 | ** | *** |
| | Multiple head shakes | 1.5 | 0.6 | **7.4** | 0.2 | **** | **** |
| | Multiple nods | 40.3 | **62.8** | 6.7 | 8.1 | **** | **** |
| | Multiple tilts | 1.2 | 1.0 | 2.2 | 1.2 | ns | ns |
| | Single head shake | 3.4 | 3.0 | 6.7 | 2.4 | ns | ns |
| | Single nod | 20.4 | **28.2** | 8.3 | 9.6 | **** | **** |
| | Single tilt | 9.3 | 7.0 | **16.0** | 10.6 | *** | **** |
| Speech | Speech | 55.1 | 70.0 | 75.6 | 7.5 | **** | ns |
| | Speech with rising F0 | 30.2 | 36.4 | **51.3** | 2.6 | **** | ** |
| | Speech with falling F0 | 24.9 | 31.8 | 33.3 | 3.3 | **** | ns |
| | Backchannel | 19.4 | 26.1 | 17.0 | 5.1 | **** | ns |
| | Not backchannel | 42.4 | 52.2 | **69.2** | 3.1 | **** | ** |
| | Speech with interrogative | 2.2 | 0.3 | **13.1** | 0.2 | **** | **** |
| | "Can't hear" | 2.9 | 0.1 | **18.3** | 0.0 | **** | **** |
| | "Can't see" | 0.3 | 0.0 | **1.9** | 0.0 | **** | **** |
| | "No" | 1.3 | 0.5 | **6.4** | 0.0 | **** | **** |
| | "Yes" | 25.4 | **41.6** | 3.5 | 0.8 | **** | **** |

*Signals are grouped by modalities. The right-most two columns present significance analyses of the distributions of the signals across clip labels, see section 4.1.1. If a signal is significantly different between positive and negative labels, the over-represented class is marked in bold.*

*†$\chi^2$ significance (Bonferroni-corrected) of the distribution on all three labels.*

*‡$\chi^2$ significance (Bonferroni-corrected) of the distribution of only positive and negative clips.*

*ns, Not significant.*

** $p < 0.05/32$.*

*** $p < 0.01/32$.*

**** $p < 0.001/32$.*

***** $p < 0.0001/32$.*

connection to a questioning tone, as mentioned in section 2.3.1. Nods and head shakes are obviously strong signals of positivity and negativity, respectively, but head jerks—the movement of the head forwards—are also correlated with negativity, or, at least, a lack of positivity, as mentioned above. Our interpretation is that this gesture generally conveys confusion and a negatively connotated surprise in our data.

**Table 2** shows that speech that is a backchannel is not significantly differently distributed between positive and negative clips, while speech that *is* a backchannel *is* significantly different between positive and negative labels. The proportion of neutral clips is higher for backchannels, and the remaining difference between positive and negative clips is small enough that the difference is not significant for backchannels. We interpret this as non-backchannel speech carrying more verbal information, usually having a more distinct meaning.

We conclude that neutral clips are associated with the *absence* of speech and head gestures, while positivity and negativity are indicated more strongly by differences between sub-labels of head movements and speech.

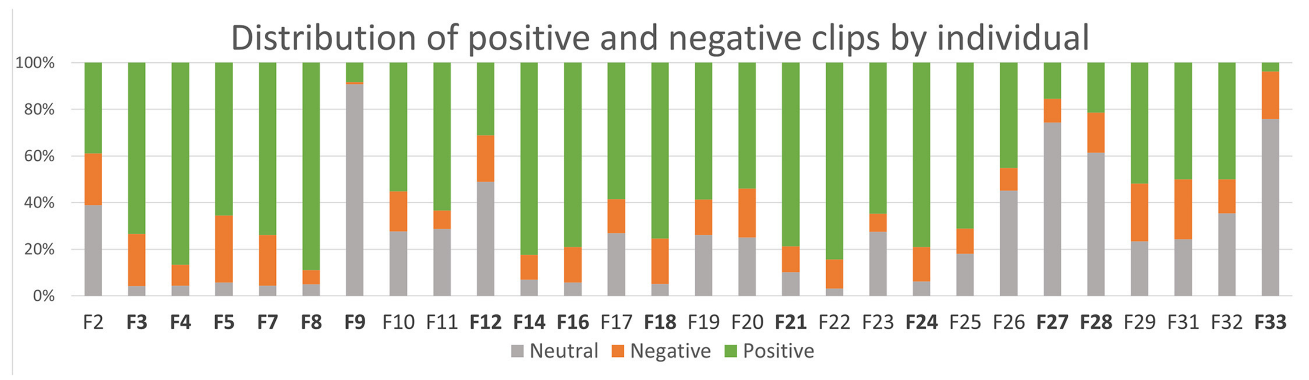

**FIGURE 2 |** Distributions of neutral, negative and positive clips for each participant. Each distribution is compared to the mean, expected distribution through a $\chi^2$ test, and individuals who significantly deviate from the mean are presented by bolding the identifier, the same as in **Figure 3**.

### 4.1.2. Individual Differences

The distribution of positive, negative, and neutral clips in the entire dataset is 59, 15, and 25%, respectively, as seen in **Table 1**. We perform a $\chi^2$ test on the participants to see if any participants deviate from the expected proportion of labels. This $\chi^2$ test has two degrees of freedom—three labels and one participant at a time—and we regard the given p-value as significant if it is lower than or equal to a Bonferroni-corrected $0.05/28 \approx 0.0018$. 15 out of the 28 participants significantly differ from the mean: the distributions and which participants are significant are presented in **Figure 2**, where the significantly different individuals are marked in bold. F9, F27, and F33 stand out by their unusually high proportion of neutral clips.

We also want to see how the usage of the multimodal signals differs between individuals. A high-level view of this distribution is to group signals by modalities. Our modality groups can be seen in the leftmost column of **Table 2**. **Figure 3** shows how the modalities used differ between the 28 individuals in our dataset. **Figure 3A** displays positive clips, and **Figure 3B** displays negative clips. As can be expected from **Table 2**, speech is slightly more common in negative clips than in positive clips, but there are outliers. Participant F27 never uses speech, in either positive or negative clips.

We perform a $\chi^2$ analysis to see if the usage of modalities differed significantly between individuals. This analysis is separated by positive, negative, and neutral clips, to find differences in how modalities are used to communicate those three labels. The $\chi^2$ test is performed on each of the 28 participants individually: it has $df = 4$, since there are five modalities, and the $\chi^2$ test has to give a p-value lower than the Bonferroni-corrected $0.05/28 \approx 0.0018$ to be considered significant. The individuals that differ significantly from the mean are denoted with bold labels in **Figure 3**. 6/28 participants differ significantly from the overall distribution for negative clips, 13/28 differ significantly for neutral clips, and 21/28 differ for positive clips. This tells us that there are significant differences in how individuals choose to use different modalities to give feedback, and that those differences are larger for positive feedback than for negative feedback. Thus, any feedback detection method relying on a single modality is likely to not work well on all subjects.

### 4.1.3. Feature Analysis Over Time

The analysis above only considers whether the signal is present or not in the clip: it does not take the timing or length of the signal into account. While **Table 2** shows that some signals are strongly associated with positivity or negativity by their very presence in a clip, it is also possible that the meaning of a signal could also depend on its timing within a clip, both in relation to the robot's speech and in relation to other signals produced by the human participant. This would not be visible in **Table 2**.

**Figure 4A** shows a tendency for positive and negative clips to have user speech after the robot finishes speaking, but negative clips have a peak later than positive clips. **Figure 4B** shows the timing of gaze on the robot, and **Figure 4C** shows the timing of gaze on the poster. These graphs show a tendency for participants in negative clips to gaze at the robot after it stops speaking. In positive clips, there is instead tendency to gaze at the poster after the robot stops speaking. Our initial hypothesis was that these patterns indicated a larger trend in the data for positively polarised signals to appear earlier in the user's turn and be shorter, and we believed that these patterns would be of the type that a timing-aware classification model would outperform one that was not timing-aware. We will come back to this hypothesis in the next section.

## 4.2. Statistical Modelling

In this section we analyse to what extent it is possible to use statistical models for predicting the three feedback labels (positive, negative, and neutral) based on the annotated features described in section 3.

### 4.2.1. Comparison of Models

Our data set is split into ten folds for use in ten fold cross-validation. The mean categorical accuracy and F-score for each model over the ten folds are presented in **Table 3**. For models that output probabilities for each class, the prediction is judged as accurate if the highest-rated class predicted by the model is

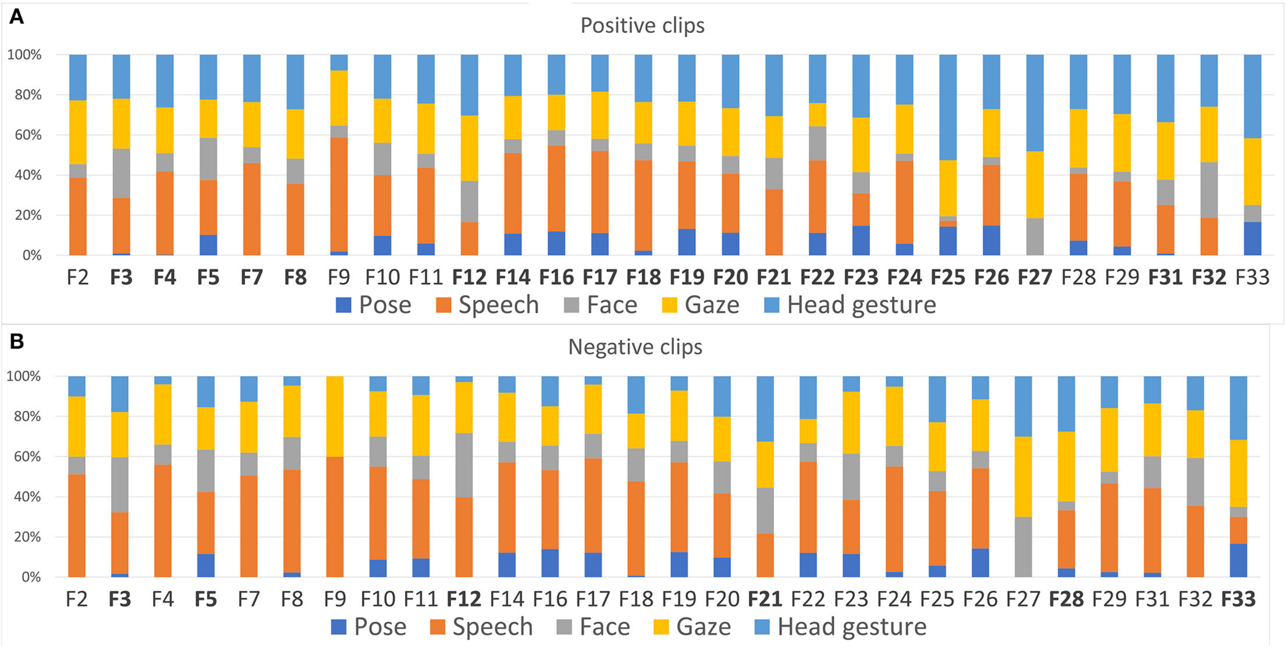

**FIGURE 3 |** The usage of different modalities of feedback from different individuals in our dataset. **(A)** Only when considering clips labelled as *positive* by the crowdworkers. **(B)** Only considering clips labelled as *negative*. In both graphs, distributions that are significantly different from the mean according to a $\chi^2$ test are marked by bolding the identifier of the participant, see section 4.1.2.

also the highest-rated class by the Mechanical Turk annotators as described in section 3.2, breaking ties in favour of neutrality over negativity over positivity.

A baseline model is introduced for comparison. This baseline model always predicts the most common clip class in the training data. This is positive clips for each fold, as can be expected from **Table 1**. The baseline model's accuracy of 59.2% therefore is exactly the proportion of positive clips in the data set as a whole. Its F-score of 0.248 is, as expected, much lower.

The two most well-performing models are the multinomial regression model on split and binary data, and the random forest model on split and fractional data. The multinomial regression model is, by the nature of what a regression model can do, not capable of handling interactions between features, like if head nods mean something different in combination with some other feature X than on their own. Compared to this, the random forest model can theoretically consider interactions between signals, but does not achieve notably higher accuracy or F-score than the multinomial regression model.

### 4.2.2. Analysis of Feature Importance

**Table 3** shows that multinomial regression and random forests perform similarly well on our data set. To evaluate which *combinations* of features had the strongest classifying power, we perform a meta-analysis with the random forest configuration that achieved the highest accuracy—with fractional and split data. For each feature in our data set, a model is trained on only that feature, and the feature that results in the model capable of the highest F-score on the training set is selected. Each combination of two features, where the first was the feature selected in the first step, is then tested in the same way, and then three features, until the F-score does not increase (on both test and training sets) upon adding a new feature. The resulting progression of features is shown in **Figure 5**. Both F-score on the training and test set are reported, but the models and features were only selected by the F-score on the training set.

The features selected by the models show a pattern where orthogonal features are selected first (*ongoing head gesture* during the participant's turn, followed by *ongoing speech* during the participant's turn). Following these, features that refine the information given by the basic features are selected. The selection of *head gesture was single tilt* as the fourth feature seems strange, but it is possible that the model selects this feature since, if it is present, it means that head gesture is less likely to be nods or head shakes, which are split across four features which may not be as important on their own.

### 4.2.3. Visualisation of RPART Trees

An advantage of the RPART models in **Table 3** is that they are easily visualised to see which features had the highest classifying power. **Figure 6** shows a tree trained on fractional and split data. The tree first splits on the presence of head gestures, and refines based on the presence of speech if there are no

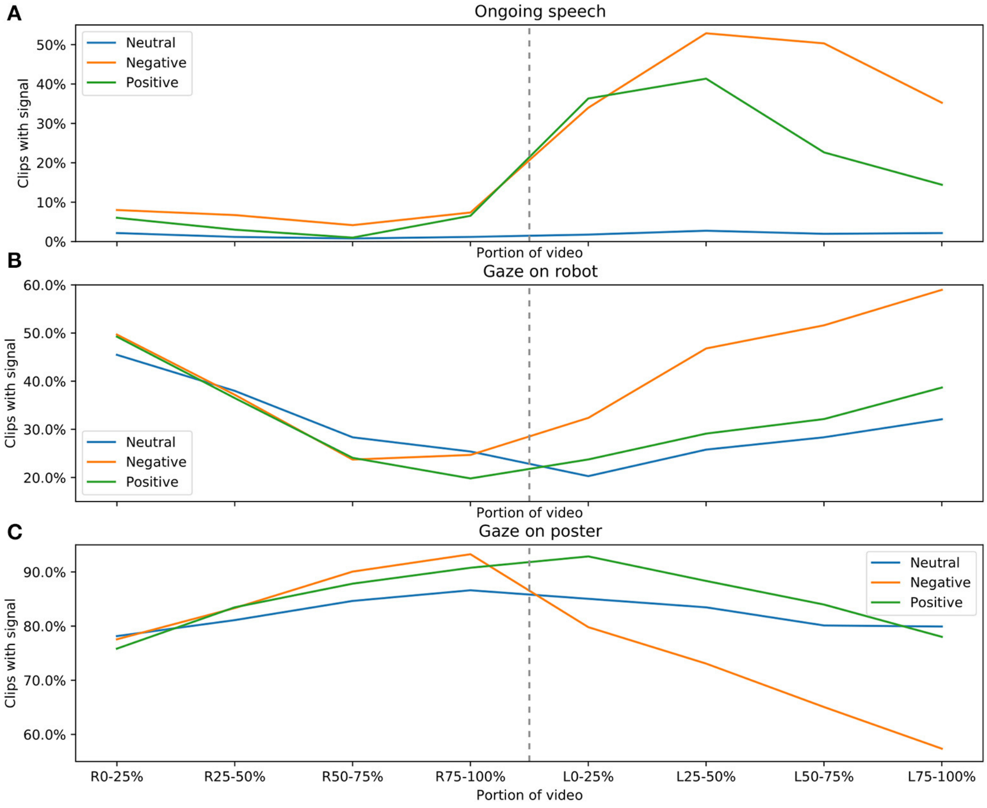

**FIGURE 4 | (A–C)** Where in the clips that the features *ongoing speech*, *gaze on robot*, and *gaze on poster*, from top to bottom, as described in section 3.3, appear. The horizontal axis has been normalised and split into two parts; on the left, the part of the clip where the robot speaks is represented with four points, representing whether the signal appears in the first 25% of the robot's speech, between 25 and 50%, between 50 and 75%, and finally for the last 25% of the part of the clip where the robot speaks. The same split is done for the part of the clip where the robot does not speak, which is similarly represented on the right half of the horizontal axis. The vertical axis shows how many of the clips that are marked as positive, negative, and neutral that contain the signal in that span. Note that all three Y axes are clipped to make the comparison between neutral, negative and positive clips clearer.

**TABLE 3 |** Accuracy and F-score for each combination of feature format and statistical model, as presented in section 3.4.

| Fractional | Split | Model | Average accuracy (%) | Average F-score |
|---|---|---|---|---|
| No | Yes | Multinomial regression | 85.857 | **0.814** |
| Yes | Yes | Random forest | **85.998** | 0.811 |
| No | Yes | Random forest | 85.372 | 0.805 |
| No | No | Random forest | 84.971 | 0.804 |
| Yes | No | Random forest | 84.755 | 0.801 |
| Yes | Yes | Multinomial regression | 84.052 | 0.797 |
| No | No | Multinomial regression | 84.414 | 0.796 |
| No | Yes | RPART tree | 85.006 | 0.795 |
| Yes | Yes | RPART tree | 84.365 | 0.785 |
| Yes | No | Multinomial regression | 82.764 | 0.783 |
| - | - | LSTM | 83.861 | 0.781 |
| No | No | RPART tree | 83.235 | 0.777 |
| Yes | No | RPART tree | 83.222 | 0.775 |
| - | - | Baseline | 59.2 | 0.248 |

*The highest values in each column have been marked in bold.*

head gestures. If there are head gestures, the tree first attempts to refine based on which head gesture was present, and falls back to classifying based on the presence of speech if this is not possible. The initial splitting by orthogonal high-level features is similar to the order found for random forests in **Figure 5**.

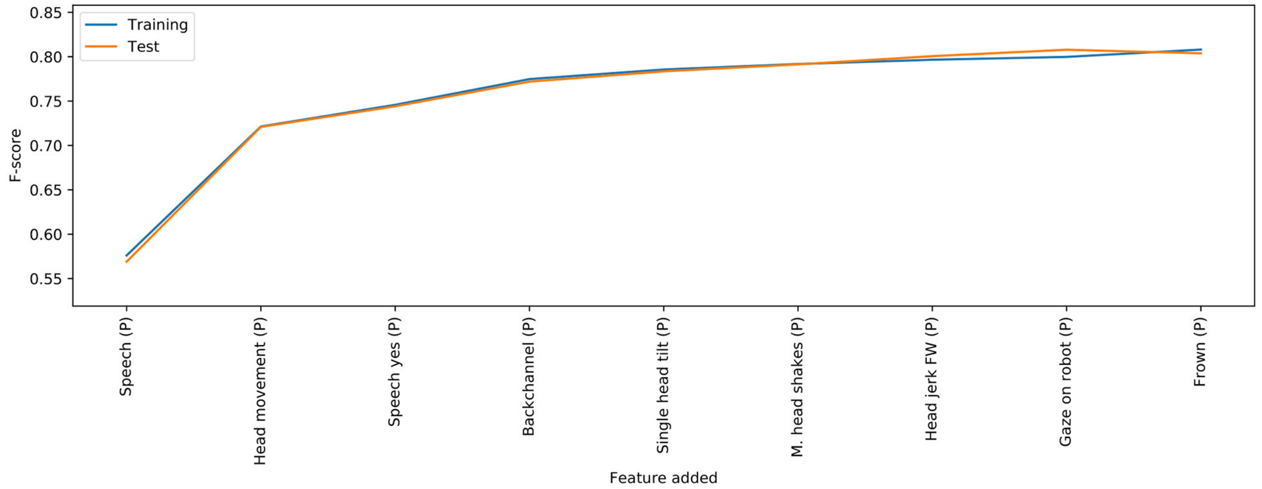

**FIGURE 5 |** The results of the feature importance analysis described in section 4.2.2, for the **random forest** model. Please note that the Y axis has been clamped.

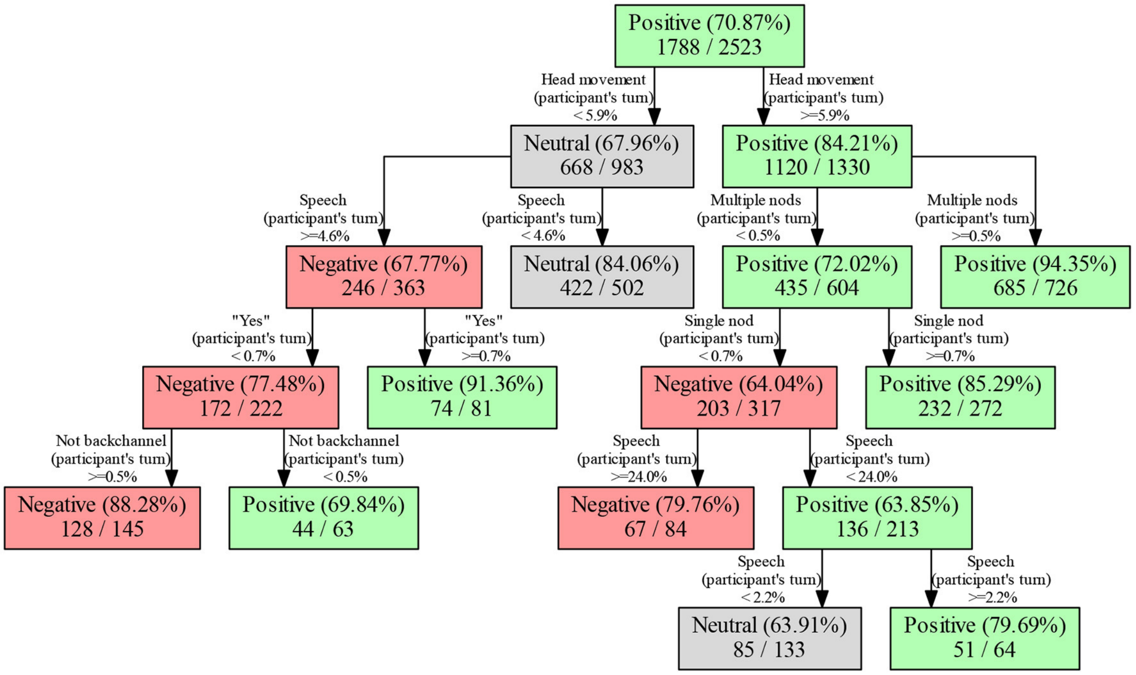
Positive (70.87%) 1788 / 2523
Head movement (participant's turn) < 5.9%
Head movement (participant's turn) >=5.9%
Neutral (67.96%) 668 / 983
Positive (84.21%) 1120 / 1330
Speech (participant's turn) >=4.6%
Speech (participant's turn) < 4.6%
Multiple nods (participant's turn) < 0.5%
Multiple nods (participant's turn) >=0.5%
Negative (67.77%) 246 / 363
Neutral (84.06%) 422 / 502
Positive (72.02%) 435 / 604
Positive (94.35%) 685 / 726
"Yes" (participant's turn) < 0.7%
"Yes" (participant's turn) >=0.7%
Single nod (participant's turn) < 0.7%
Single nod (participant's turn) >=0.7%
Negative (77.48%) 172 / 222
Positive (91.36%) 74 / 81
Negative (64.04%) 203 / 317
Positive (85.29%) 232 / 272
Not backchannel (participant's turn) >=0.5%
Not backchannel (participant's turn) < 0.5%
Speech (participant's turn) >=24.0%
Speech (participant's turn) < 24.0%
Negative (88.28%) 128 / 145
Positive (69.84%) 44 / 63
Negative (79.76%) 67 / 84
Positive (63.85%) 136 / 213
Speech (participant's turn) < 2.2%
Speech (participant's turn) >=2.2%
Neutral (63.91%) 85 / 133
Positive (79.69%) 51 / 64

**FIGURE 6 |** The RPART tree for fractional, split data, with fold 6 used as test data.

### 4.2.4. Muting Modalities

The individual modalities shown in **Table 2**—pose, face, gaze, head gestures, and speech—refer to separate, possibly co-occurring ways to send feedback signals. To explore which of the modalities were less important, and which modalities could be expressed through combinations of other modalities, we train a random forest model on every combination of including and not including each modality. The results are presented in **Table 4**.

If including a certain modality would lead to overfitting, we could expect the model to perform better when excluding that modality. As can be seen in **Table 4**, including every modality does not lead to overfitting—at least for a random forest model—and there is a logical binary pattern where removing pose has the smallest effect, followed by gaze, face, head gestures, and speech in order. The model that is trained on data without any multimodal features performs worse than the baseline. We

**TABLE 4 |** An ordered presentation of F-scores and accuracies when random forest models are not given certain modalities.

| Speech | Head | Face | Gaze | Pose | Average accuracy (%) | Average F-score (%) |
|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | ✓ | 85.992 | 81.233 |
| ✓ | ✓ | ✓ | ✓ |  | 85.934 | 81.078 |
| ✓ | ✓ | ✓ |  | ✓ | 85.648 | 80.858 |
| ✓ | ✓ | ✓ |  |  | 85.67 | 80.801 |
| ✓ | ✓ |  | ✓ | ✓ | 85.755 | 80.726 |
| ✓ | ✓ |  | ✓ |  | 85.637 | 80.696 |
| ✓ | ✓ |  |  | ✓ | 85.552 | 80.54 |
| ✓ | ✓ |  |  |  | 85.492 | 80.425 |
| ✓ |  | ✓ |  | ✓ | 76.332 | 73.329 |
| ✓ |  | ✓ |  |  | 75.06 | 72.065 |
| ✓ |  | ✓ | ✓ |  | 74.59 | 71.668 |
| ✓ |  | ✓ | ✓ | ✓ | 74.951 | 71.101 |
| ✓ |  |  |  | ✓ | 72.088 | 69.749 |
| ✓ |  |  |  |  | 71.704 | 69.719 |
| ✓ |  |  | ✓ | ✓ | 73.002 | 69.499 |
| ✓ |  |  | ✓ |  | 72.317 | 69.442 |
|  | ✓ | ✓ | ✓ | ✓ | 75.663 | 66.53 |
|  | ✓ | ✓ | ✓ |  | 75.325 | 66.138 |
|  | ✓ |  | ✓ |  | 74.574 | 65.635 |
|  | ✓ |  | ✓ | ✓ | 74.434 | 65.375 |
|  | ✓ | ✓ |  | ✓ | 74.617 | 64.197 |
|  | ✓ | ✓ |  |  | 74.335 | 63.347 |
|  | ✓ |  |  | ✓ | 72.392 | 57.376 |
|  | ✓ |  |  |  | 72.067 | 56.641 |
|  |  | ✓ | ✓ | ✓ | 62.698 | 44.189 |
|  |  | ✓ | ✓ |  | 61.653 | 42.395 |
|  |  | ✓ |  |  | 59.695 | 35.673 |
|  |  | ✓ |  | ✓ | 60.569 | 35.137 |
|  |  |  | ✓ | ✓ | 59.8 | 34.473 |
|  |  |  | ✓ |  | 59.152 | 33.541 |
|  |  |  |  | ✓ | 59.107 | 24.945 |
|  |  |  |  |  | 59.199 | 24.771 |

*This table represents specifically the **random forest, fractional, split data** model which achieves the highest accuracy in **Table 3**.*

interpret this as a result of overfitting on the *is elicitation* meta-feature that remained.

### 4.2.5. Cross-Validation Leaving One Participant Out at a Time

The 10-fold validation we use for our main statistical model evaluation has the advantage of ensuring that test data and training data have comparable distributions of positive, negative, and neutral clips. This is not the case if the choices for training and test data are made on a participant-by-participant basis. However, leaving a single participant at a time out of the training data does illustrate if their behaviours are similar to or different to the behaviours expressed by the other participants left in the training data. It is also a better showcase of whether the trained models generalise to behaviours from individuals they have never seen before. In light of this argument, we perform 28-wise cross-validation on our dataset, for specifically the random forest model that achieves the highest accuracy in **Table 3**. Each participant is used for test data exactly once, with the others being used for training data. The results are presented in **Figure 7**. Each participant is represented by the F-score and accuracy of the model where they were the test set.

To analyse the results in **Figure 7**, we want to know if participants with notably lower accuracy and/or F-score in **Figure 7** correspond to the bolded participants in **Figure 3**. Looking at the three participants with the lowest F-score—F3, F27, and F33—we see that F3 has significantly deviating usage of modalities in both positive and negative clips, F27 deviates from the mean only for positive clips, and F33 only for negative clips. Both F27 and F33 have unusually high proportions of neutral clips (see section 4.1.2), so the model's difficulty in estimating their polarity based on the training data is presumably because of their high rate of neutral clips, corresponding to a low rate of feedback signals overall. Notably, however, F9, who has the highest rate of neutral clips overall with 88/97 (see **Figure 2**), is one of the participants whose model gets the highest accuracy in

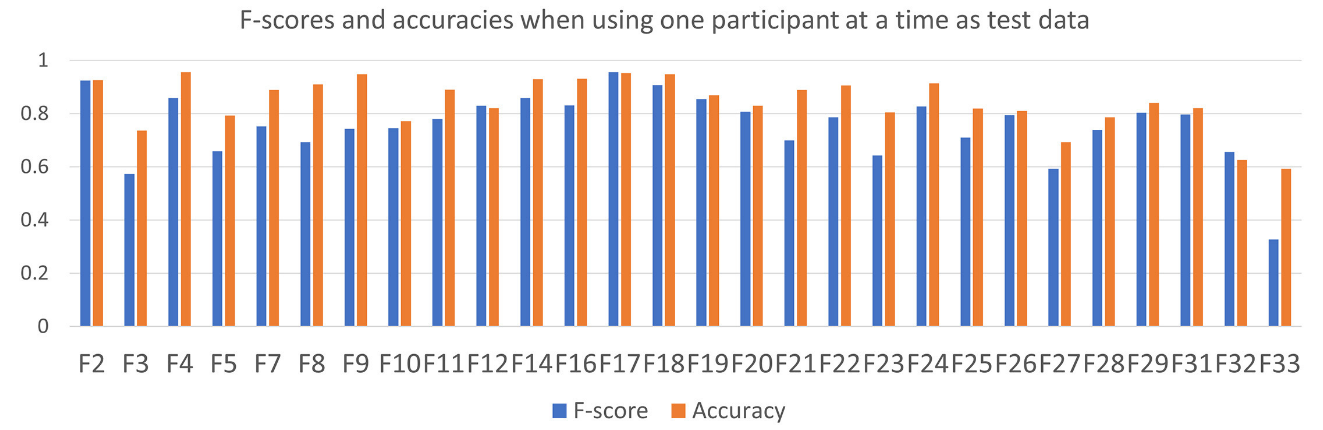

**FIGURE 7 |** F-score (blue, the left bar of each pair) and accuracy (orange, the right bar of each pair) when training a random forest model on split, fractional data. For each of the 28 models, one individual is left out as test data—this is the individual pointed out on the X axis label. The rest of the dataset is used as training data. This figure thus shows how certain individuals use behaviours that are harder to interpret based on training on the behaviours of the other individuals.

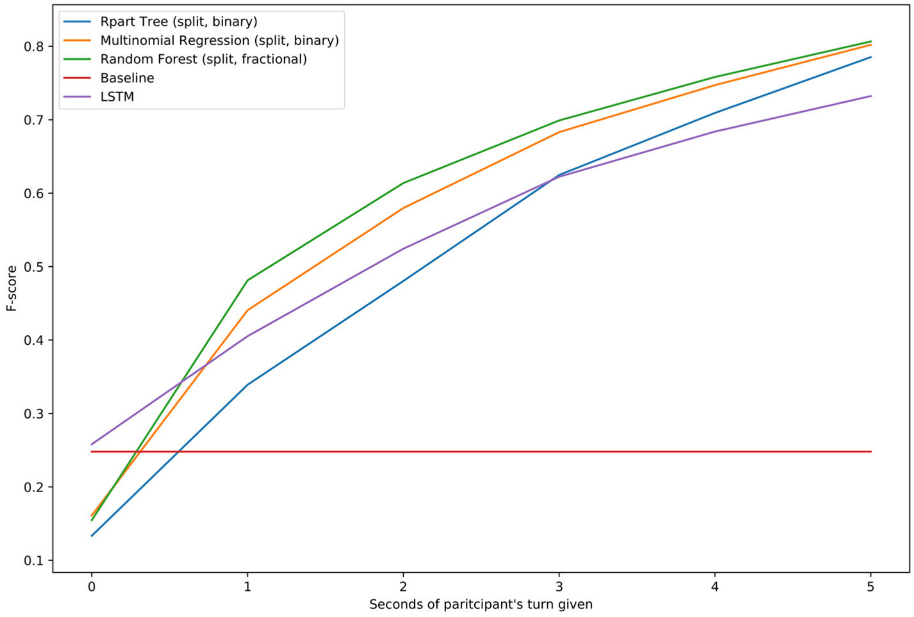

**FIGURE 8 |** The F-score of the various models from **Table 3** when they are only given data corresponding to the first second of the participant's turn, first 2 s of the participant's turn, and so on until 5 s, at which point the F-scores stop going up or down, even though there are a small number of clips with a participant turn as long as 11 s.

**Figure 7**. Clearly, for F9, the model is able to generalise that an absence of signals is a sign of neutrality, but this strategy does not work for F33 and F27. F27 and F33 have a lack of speech signals in common—F27 does not speak at all—while F9 is very active in the speech modality but uses no head gestures.

### 4.2.6. Model Evaluations Over Time

The LSTM model presented in section 3.4.4 operates on data given to it in time-frames representing a tenth of a second. While the other statistical models in **Table 3** are not inherently time-aware in this way, they can still be trained and used to give an evaluation over time by passing them data reflecting what has happened up to a point in the clip.

**Figure 8** reports the F-score of models when given data corresponding to the first second of each participant's turn, the first 2 s, and so on. The X axis has been clipped at 5 s, since the values converge at this point. Five seconds corresponds to the end of the participant's turn in most of our clips. Notably, the LSTM model is slightly better than the other statistical models right at the start of the user's turn; we assume this is because the LSTM model has been able to use the timing and presence of signals during the robot's turn into account to create a slightly better assumption of what the user's reaction is going to be. As more time passes, however, all statistical models eventually surpass the LSTM in both F-score and accuracy (not shown), around 2 s in. The models other than the LSTM are only trained on data corresponding to the user's full turn, so the fact that they outperform the LSTM at almost all points in time is a strong indication that the polarity of a clip is mostly defined by the presence of signals, regardless of their timing.

## 5. DISCUSSION

We will now return back to the questions posed in section 1 and try to answer them in light of the findings from this study.

**1. What modalities are most commonly used to convey negative and positive feedback?**

**Table 2** shows that head-nods (multiples or single nods) are the strongest indicator of positive feedback, whereas head shakes and tilts indicate negative feedback. When it comes to facial expressions, the only clear signal is frowning, which indicates negative feedback. Non-backchannel speech is most often used to express negative feedback, whereas backchannels can be both negative and positive. Rising F0 is also associated with negative feedback. **Table 2** does not tell us which signals are a strong indication of a neutral clip. However, the model analyses we have presented in section 4.2.1 suggest that the strongest indication of a neutral clip is an absence of any strong signals for either positive or negative feedback.

By comparison to the above, there are signals in our dataset that we would have expected to be connected to certain polarities, but which show no such significance. Shrugging is too rare to be a sign of anything, but if it were more common, **Table 2** suggests that it would be a sign of negativity, or at least non-neutrality. Eyebrow raises are not, as we would expect, a sign of negativity, but appear relatively commonly in positively labelled clips as well, indicating that surprise is as often positive as it is negative.

Our scenario and experiment set-up may have affected which signals users tended to use. The turn-taking heuristic we used defaulted to a turn-time of 5 s—if the user had not reacted with feedback that could be classified by the Wizard of Oz within this time, the system would produce an elicitation. The Wizard had the capacity to shorten or lengthen the user's turn in response to feedback where this felt natural, but we can see from **Figure 8** that the models reach their maximum performance after 5 s. Even though our system had the capacity to allow for user turns shorter or longer than 5 s, it appears that users generally synchronised with its preferred cadence of 5-s turns. This cadence of feedback presumably restricted users from reacting with longer speech and sequences of feedback, even when they would have liked to do so. On the other hand, this restriction is not entirely inappropriate for our museum guide scenario—a museum guide does not necessarily want their audience to constantly interrupt their presentation, depending on how scripted and prepared the presentation is.

Therefore, we believe that **Tables 2**, **4** accurately depict which modalities and groups of modalities are most appropriate to pick up for the scenario of a presentation agent, but further studies need to be done to find out whether this would also be true for other scenarios—where the robot is a more conversing, less driving agent. The relative unimportance of hand gestures from our listeners also matches up with earlier results from Battersby (2011).

The results of Kuno et al. (2007) suggested that nods and gaze were important signals of a user being involved with a presenting robot's presentation. While their results match with ours when it comes to head-nods, gaze at first appears to have been more important for their participants than ours. However, looking at **Figure 4**, participants did in fact generally gaze on the poster along with the robot, regardless of if the clip was positive, negative, or neutral. This feature may not be unimportant for determining if a participant is engaged in a presentation, but since both positively and negatively classified clips assume that the participant is engaged, the difference in importance is not necessarily a disagreement in results.

Oppenheim et al. (2021) showed that the feedback responses used by test participants were significantly different depending on if the speaker gazed at the listener to *supplement*, *highlight*, or *converse*, with speech being less common than nods, as common as nods, and more common than nods, respectively. Our presenting robot agent predictably gazed at the listener at the end of each line. We see that nods, single or multiple, appear in more clips than speech, by the frequencies in **Table 2**. Our robot's motivation for gaze was always closer to the *supplement* label by Oppenheim et al. (2021) than the other two, since our robot had finished speaking by the time it gazed at the user, and never intended to hand over the turn for more than a brief comment. Thus, our results roughly correspond to the proportions seen by Oppenheim et al. (2021).

Rising F0 is an indication of negativity in our dataset, as seen in **Table 2**. Because of the relatively short user turns, and because user turns were restricted to being feedback or *track 2* comments on the content presented by the robot, we can presume that prosody was not used to invite backchannels or highlight given or non-given information, as mentioned in section 2.3.1. This leaves the use of prosody to mark a proposition as a question, or to ask to receive more information about some aspect about the information previously presented, as mentioned by Hirschberg and Ward (1995) and Bartels (1997). It is possible to use this type of prosody to mark a question that we would have labelled as positivity ("And when was that?", "Why did that happen?"), but one interpretation of our data is that participants to a large extent used such questions to ask the robot to repeat itself, or explain something they had not understood.

Like Malisz et al. (2016), we also found that nods more commonly occurred in groups than one-by-one, see **Table 2**. Malisz et al. (2016) also found the same pattern for head shakes, which we do not significantly see in our corpus. This could be because Malisz et al. (2016) see proportionally much fewer head-shakes than we do, only registering 35 head-shakes in a corpus also containing 1,032 nods—a completely different ratio than ours, and hard to compare because we specifically elicited negative feedback from our participants by making the robot presenter misspeak. Like Malisz et al. (2016), however, we also see that single tilts are more common than multiple tilts.

**2. Are any modalities redundant or complimentary when it comes to expressing positive and negative feedback?**

**Table 4** tells us that Speech and Head are the most important modality groups and when only using these two modalities, the F-score is quite close to using all modalities (80.4 vs. 81.2%). Thus, even though **Table 2** showed that frowning was associated with negative feedback, Face, Gaze, and Pose do not have much overall impact on the classification of feedback type and can be considered fairly redundant. **Table 4** also shows that when only using Speech or Head on their own, the performance drops significantly (69.7 and 56.6%). Thus, they seem to be highly complimentary to each other.

**3. Does the interpretation of feedback as positive or negative change based on its relative timing to other feedback and the statement being reacted to?**

We were expecting the interpretation and ordering of feedback in our model to affect the meaning in terms of positivity or negativity, but this does not seem to hold based on the results we have presented. Models which are simply given the presence of a signal, ignoring internal order and timing, perform better on classifying our dataset as positive, negative, or neutral than the timing-aware LSTM model. The three most high-performing models in **Table 3** are split models—meaning that they received data that differentiated between if a signal was used during the robot's turn or the participant's response. This indicates that multinomial regression and random forest models benefit from the distinction between these timings, and that some information is contained in it. However, the timing within the user's turn does not appear to matter, or at least matters much less than the identity of the signal.

**Figures 5**, **6** show that the features that describe whether a user used a signal during their turn, after the robot stopped speaking, carry more information than the signals from when the robot was speaking. In fact, in both **Figures 5**, **6**, the *only* features that appear are signals denoting the user's turn. This tells us that the relative performance advantage of *split* models in **Table 3** is because they were able to ignore what the user did during the robot's turn.

An important question is **why** there do not seem to be timing and ordering effects in our dataset. One explanation is that the scenario—passive audiences to a museum guide presenting facts about a painting—lends itself to the audience delivering one strong piece of positive feedback when prompted. It is also possible that our agent design prompted this type of behaviour in its audience because of the turn-taking cadence and elicitation patterns. It has been previously established that users use mostly the modalities and signals that they expect a system like ours to recognise (Sidner et al., 2006; Kontogiorgos et al., 2021; Laban et al., 2021).

Another potential explanation of the relative unimportance of timing and ordering is that those effects *are* present in our data, but are not necessary for predicting our positive/negative/neutral labels—they could, however, be useful for a more in-depth grounding annotation, using labels similar to those by Clark (1996).

**4. Are there individual differences in the use of modalities to communicate different polarities of feedback?**

As reported in section 4.1.2, many participants had significantly differing distributions (from the mean) of modalities used for expressing negative and positive feedback. **Figure 3** illustrates these differences. Speech appears to have been the dominant way to express negative feedback. Positive feedback is expressed with signals that are split between head gestures and speech, especially the "yes" signal, as seen in **Table 2**. Since positive clips are more common than negative or neutral clips in our dataset, it is also not surprising that participants are able to use a larger variety of signals in those clips. We have been unable to find previous literature that describes if humans generally use more varied feedback to express positive feedback than negative feedback.

Speech and head movements are not strictly positive or negative modalities—but sub-signals within the modality can be significantly positive or negative, as shown in **Table 2**. Head nods and head shakes are unsurprisingly positive and negative, respectively, in our dataset: "yes" and "no" can be seen as the spoken counterparts of these signals, and are similarly significantly positive and negative. These signals can be seen as encoding attitudinal reactions to the content spoken by the robot—they only have a meaning if the user understood what the robot was saying.

**Figure 7** and the arguments presented in section 4.1.2 indicate that the hardest individuals to classify based on training on the other individuals in our dataset are the ones that are disproportionately labelled as neutral because they do not use many feedback signals. Participants like F9, who use feedback in an ambiguous way, are easier to classify as neutral. The problem for our models is not classifying feedback as positive and negative, but rather what to do when that feedback is not present. The neutral label is more common than the negative label in the dataset, so by the numbers, correctly classifying participants as neutral is more important than being able to classify them as negative.

Navarretta et al. (2012) showed that Finnish participants used single nods more than multiple nods. In our dataset, multiple nods are significantly more common than single nods. This could be explained by many of the participants being Swedish, as Navarretta et al. (2012) showed that Swedish and Danish subjects preferred multiple nods to single nods—and even for those participants who were not native Swedish speakers, it could be argued that they were using feedback patterns similar to the Swedish environment in which they live. The corpus study by Malisz et al. (2016) showed that multiple nods were also more common in a German-speaking context, and since most of our participants were Western European, the fact that multiple nods were more common than single nods could be a sign of a regional pattern where Western Europe prefers multiple nods to single nods. Nonetheless, both single and multiple nods were positive signals, so individual differences in which of the two signals an individual chooses to use would not complicate feedback classification.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an analysis of how humans express negative and positive feedback across different modalities when listening to a robot presenting a piece of art. The results show that the most important information can be found in their speech and head movements (e.g., nods), whereas facial expressions, gaze direction and body pose did not contribute as much. There seems to be more variation between individuals when it comes to how positive feedback is expressed, compared to negative feedback. Often, the very presence of a nod, a head shake, or certain speech is enough to classify an entire reaction as a positive or negative reaction, regardless of the context. The precise timing of the feedback does not seem to be of importance.

For future research, we note that our analyses of the gaze and pose modalities were not as deep as the analyses of the speech and head modalities. An interesting direction for future work in feedback analysis for presentation agents could be to enhance gaze and pose analysis with more detailed sub-signals, like hand gesture sensing and more detailed approximations of gaze targets. We have shown that not much positive or negative information is contained in whether the participant looks at the presented object or the presenting agent, but it is still possible that gaze sub-targets within the presented objects carry information that we were not able to annotate or extract.

We were unable to annotate our dataset with a rich grounding scheme like that of Clark (1996), and fell back on the labels positive/negative/neutral. It's possible that annotating the data with employed professional annotators would have led to the more in-depth annotation succeeding, like in the work by Malisz et al. (2016). While we did not see the ordering and timing effects that we were expecting to see—see Question 3 in section 5— it is possible that such effects come into play when the models are asked to perform a more fine-grained classification with four grounding levels, rather than the simpler positive/negative/neutral labels. One advantage of the rich multimodal annotation of our dataset is that many of the signals listed in **Table 2** carry strong implications about what grounding level the classified feedback must be on—if our statistical models report that a clip is positive, and the "yes" feature is present in the clip, for example, we can conclude that the feedback must at least mean *understanding*, if not *acceptance*. This allows us to partially reconstruct grounding data akin to the standards by Clark (1996) and Allwood et al. (1992) from our simpler classification.

The results are important for the development of future adaptive presentation agents (which could be museum guides or teachers), as they indicate that such systems should focus on the analysis of speech and head movements, and put less focus on the analysis of the audience's facial expressions, gaze or pose. The results indicate that such an agent should be able to determine fairly reliably whether user feedback is positive, negative, or neutral. If positive, the presentation can proceed, and if negative, the agent can try to repair or reformulate the presentation. If only neutral (i.e., absence of) feedback is received for too long, the agent should elicit (positive or negative) feedback from the user (depending on the *grounding criterion*, as discussed in section 2.2). An example of such a framework, where this kind of classification would be of direct use, is the model we have presented in Axelsson and Skantze (2020).

## DATA AVAILABILITY STATEMENT

The raw ELAN files as well as data files used to train the statistical models in this paper are published with DOI 10.5281/zenodo.5078235, accessible at https://doi.org/10.5281/zenodo.5078235.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

AA collected and analysed the data from the experiments, programmed the statistical models, and analysed their behaviours. GS supervises AA and co-wrote the paper. Both authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the SSF (Swedish Foundation for Strategic Research) project Co-adaptive Human-Robot Interactive Systems (COIN).

## ACKNOWLEDGMENTS

We would like to thank the annotators of our dataset and the reviewers for their insightful comments.

## REFERENCES

Agarwal, A., Yadav, A., and Vishwakarma, D. K. (2019). "Multimodal sentiment analysis via RNN variants," in *Proceedings - 2019 IEEE/ACIS 4th International Conference on Big Data, Cloud Computing, and Data Science, BCD 2019* (Honolulu, HI), 19–23. doi: 10.1109/BCD.2019.8885108

Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). "FurHat: a back-projected human-like robot head for multiparty human-machine interaction," in *Lecture Notes in Computer Science* (Berlin: Springer), 114–130. doi: 10.1007/978-3-642-34584-5_9

Allwood, J., Cerrato, L., and Dybkjaer, L. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Lang. Resour. Eval.* 41, 273–287. doi: 10.1007/s10579-007-9061-5

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, N., and Skantze, G. (2019). "Modelling adaptive presentations in human-robot interaction using behaviour trees," in *SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse 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 ACM International Conference on Intelligent Virtual Agents, IVA 2020* (Glasgow), 1–8. doi: 10.1145/3383652.3423884

Baker, M., Hansen, T., Joiner, R., and Traum, D. (1999). The role of grounding in collaborative learning tasks. *Collab. Learn. Cogn. Comput. Approch*. 31, 31–63.

Bartels, C. (1997). "The pragmatics of Wh-question intonation in English," in *University of Pennsylvania Working Papers in Linguistics* (Philadelphia, PA), 1–17.

Battersby, S. A. (2011). *Moving together: the organisation of non-verbal cues during multiparty conversation* (Ph.D. thesis). Queen Mary University of London, London, United Kingdom.

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

Bertrand, R., Ferré, G., Blache, P., Espesser, R., and Rauzy, S. (2007). "Backchannels revisited from a multimodal perspective," in *International Conference on Auditory-Visual Speech Processing 2007 (AVSP2007)* (Hilvarenbeek: ISCA). Available online at: https://hal.archives-ouvertes.fr/hal-00244490/document

Boersma, P., and van Heuven, V. (2001). Speak and unspeak with Praat. *Glot Int*. 5, 341–347. Available online at: https://www.fon.hum.uva.nl/paul/praat.html

Bohus, D., and Rudnicky, A. (2006). "A "K hypotheses + other" belief updating model," in *AAAI Workshop - Technical Report* (Menlo Park, CA), 13–18.

Buck, R. (1980). Nonverbal behavior and the theory of emotion: the facial feedback hypothesis. *J. Pers. Soc. Psychol*. 38, 811–824. doi: 10.1037/0022-3514.38.5.811

Buschmeier, H., and Kopp, S. (2011). "Unveiling the information state with a Bayesian model of the listener," in *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue* (Los Angeles, CA), 178–179.

Buschmeier, H., and Kopp, S. (2013). Co-constructing grounded symbols-feedback and incremental adaptation in human-agent dialogue. *Künstliche Intelligenz* 27, 137–143. doi: 10.1007/s13218-013-0241-8

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

Clark, H. H. (1994). Managing problems in speaking. *Speech Commun*. 15, 243–250. doi: 10.1016/0167-6393(94)90075-2

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

Clark, H. H., and Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on Socially Shared Cognition*, eds L. B. Resnick, J. M. Levine, and S. D. Teasley (Pittsburgh, PT: American Psychological Association), 127–149. doi: 10.1037/10096-006

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

Colletta, J. M., Pellenq, C., and Guidetti, M. (2010). Age-related changes in co-speech gesture and narrative: evidence from French children and adults. *Speech Commun*. 52, 565–576. doi: 10.1016/j.specom.2010.02.009

Duncan, S., and Fiske, D. W. (1977). *Face-to-Face Interaction: Research, Methods, and Theory*. Oxfordshire: Routledge.

Edinger, J. A., and Patterson, M. L. (1983). Nonverbal involvement and social control. *Psychol. Bull*. 93, 30–56. doi: 10.1037/0033-2909.93.1.30

Edlund, J., House, D., and Skantze, G. (2005). "The effects of prosodic features on the interpretation of clarification ellipses," in *9th European Conference on Speech Communication and Technology* (Lisbon), 2389–2392. doi: 10.21437/Interspeech.2005-43

Ekman, P. (2004). "Emotional and conversational nonverbal signals," in *Language, Knowledge, and Representation*, eds J. M. Larrazabal and L. A. P. Miranda (Berlin: Springer), 39–50. doi: 10.1007/978-1-4020-2783-3_3

Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., et al. (2021). Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educ. Psychol. Rev*. 33, 27–49. doi: 10.1007/s10648-019-09514-z

Goswami, M., Manuja, M., and Leekha, M. (2020). Towards social & engaging peer learning: predicting backchanneling and disengagement in children. *arXiv [preprint]. arXiv:2007.11346*.

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

Heylen, D. (2005). "Challenges ahead head movements and other social acts in conversations," in *AISB'05 Convention: Proceedings of the Joint Symposium on Virtual Social Agents: Social Presence Cues for Virtual Humanoids Empathic Interaction with Synthetic Characters Mind Minding Agents* (Hertfordshire), 45–52.

Hirschberg, J., and Ward, G. (1995). The interpretation of the high-rise question contour in English. *J. Pragmat*. 24, 407–412. doi: 10.1016/0378-2166(94)00056-K

Hirst, D., and Di Cristo, A. (1998). "A survey of intonation systems," in *Intonation Systems: A Survey of Twenty Languages*, eds D. Hirst, and A. Di Cristo (Cambridge: Cambridge University Press), 1–44.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat*. 15, 651–674. doi: 10.1198/106186006X133933

Hough, J., and Schlangen, D. (2017). "It's not what you do, it's how you do it: grounding uncertainty for a simple robot," in *ACM/IEEE International Conference on Human-Robot Interaction* (Vienna), 274–282. doi: 10.1145/2909824.3020214

Hsieh, W. F., Li, Y., Kasano, E., Simokawara, E. S., and Yamaguchi, T. (2019). "Confidence identification based on the combination of verbal and non-verbal factors in human robot interaction," in *Proceedings of the International Joint Conference on Neural Networks* (Budapest), 1–7. doi: 10.1109/IJCNN.2019.8851845

Iio, T., Satake, S., Kanda, T., Hayashi, K., Ferreri, F., and Hagita, N. (2020). Human-like guide robot that proactively explains exhibits. *Int. J. Soc. Robot*. 12, 549–566. doi: 10.1007/s12369-019-00587-y

Jain, V., Leekha, M., Shah, R. R., and Shukla, J. (2021). Exploring semi-supervised learning for predicting listener backchannels. *arXiv preprint arXiv:2101.01899*. doi: 10.1145/3411764.3445449

Jimenez-Molina, A., Retamal, C., and Lira, H. (2018). Using psychophysiological sensors to assess mental workload during web browsing. *Sensors* 18:458. doi: 10.3390/s18020458

Jokinen, K. (2009). "Nonverbal feedback in interactions," in *Affective Information Processing*, eds J. Tao and T. Tan (Berlin: Springer), 227–240. doi: 10.1007/978-1-84800-306-4_13

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. G. Barres, Z. C. Carrion, and R. D. Lopez-Cozar (Pennsylvania, PN: IGI Global), 38–58. doi: 10.4018/978-1-4666-2530-3.ch003

Kleinke, C. L. (1986). Gaze and eye contact. A research review. *Psychol. Bull*. 100, 78–100. doi: 10.1037/0033-2909.100.1.78

Kontogiorgos, D., Pereira, A., and Gustafson, J. (2019). "Estimating uncertainty in task-oriented dialogue," in *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction* (Suzhou), 414–418. doi: 10.1145/3340555.3353722

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

Krahmer, E., Swerts, M., Theune, M., and Weegels, M. (2002). The dual of denial: two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Commun*. 36, 133–145. doi: 10.1016/S0167-6393(01)00030-9

Krauss, R. M., Chen, Y., and Chawla, P. (1996). Nonverbal behavior and nonverbal communication: what do conversational hand gestures tell us? *Adv. Exp. Soc. Psychol*. 28, 389–450. doi: 10.1016/S0065-2601(08)60241-5

Kuno, Y., Sadazuka, K., Kawashima, M., Yamazaki, K., Yamazaki, A., and Kuzuoka, H. (2007). "Museum guide robot based on sociological interaction analysis," in *Conference on Human Factors in Computing Systems - Proceedings* (San Jose, CA), 1191–1194. doi: 10.1145/1240624.1240804

Kuzuoka, H., Suzuki, Y., Yamashita, J., and Yamazaki, K. (2010). "Reconfiguring spatial formation arrangement by robot body orientation," in *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010* (Osaka), 285–292. doi: 10.1109/HRI.2010.5453182

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

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33:159. doi: 10.2307/2529310

Ma, J., Zheng, W. L., Tang, H., and Lu, B. L. (2019). "Emotion recognition using multimodal residual LSTM network," in *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia* (Nice), 176–183. doi: 10.1145/3343031.3350871

Malisz, Z., Włodarczak, M., Buschmeier, H., Kopp, S., and Wagner, P. (2012). "Prosodic characteristics of feedback expressions in distracted and non-distracted listeners," in *Proceedings of The Listening Talker. An Interdisciplinary Workshop on Natural and Synthetic Modification of Speech in Response to Listening Conditions* (Edinburgh), 36–39.

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

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

Mehlmann, G., Janowski, K., and André, E. (2016). Modeling grounding for interactive social companions. *Künstliche Intelligenz* 30, 45–52. doi: 10.1007/s13218-015-0397-5

Mehlmann, G., Janowski, K., Häring, M., Baur, T., Gebhard, P., and André, E. (2014). "Exploring a model of gaze for grounding in multimodal HRI," in *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction* (Istanbul), 247–254. doi: 10.1145/2663204.2663275

Nakano, Y. I., and Ishii, R. (2010). "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *International Conference on Intelligent User Interfaces, Proceedings IUI* (Hong Kong), 139–148. doi: 10.1145/1719970.1719990

Nakatsukasa, K., and Loewen, S. (2020). "Non-verbal feedback," in *Corrective Feedback in Second Language Teaching and Learning*, eds H. Nassaji and E. Kartchava (New York, NY: Routledge), 158–173. doi: 10.4324/9781315621432-12

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., and Paggio, P. (2012). "Feedback in nordic first-encounters: a comparative study," in *LREC*, eds N. Calzolari,, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, et al (Istanbul: European Language Resources Association), 2494–2499.

Nourbakhsh, I. R., Kunz, C., and Willeke, T. (2003). "The mobot museum robot installations: a five year experiment," in *IEEE International Conference on Intelligent Robots and Systems* (Las Vegas, NV), 3636–3641. doi: 10.1109/IROS.2003.1249720

Novick, D. (2012). "Paralinguistic behaviors in dialog as a continuous process," in *Interdisciplinary Workshop on Feedback Behaviors in Dialog* (Stevenson, WA).

Novick, D., and Gris, I. (2013). "Grounding and turn-taking in multimodal multiparty conversation," in *Lecture Notes in Computer Science*, ed K. Masaaki (Berlin: Springer), 97–106. doi: 10.1007/978-3-642-39330-3_11

Oertel, C., Lopes, J., Yu, Y., Mora, K. A. F., 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

Oppenheim, J., Huang, J., Won, I., and Huang, C. M. (2021). "Mental synchronization in human task demonstration: implications for robot teaching and learning," in *ACM/IEEE International Conference on Human-Robot Interaction* (Virtual Conference), 470–474. doi: 10.1145/3434074.3447216

Paek, T., and Horvitz, E. (2000). Grounding criterion: toward a formal theory of grounding.

Park, H. W., Grover, I., Spaulding, S., Gomez, L., and Breazeal, C. (2019). "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (Honolulu, HI), 687–694. doi: 10.1609/aaai.v33i01.3301687

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005). "Engagement capabilities for ECAs," in *AAMAS'05 Workshop: Creating Bonds with ECAs*, Utrecht.

Pichl, J., Marek, P., Konrád, J., Matulík, M., and Šedivý, J. (2020). Alquist 2.0: Alexa prize socialbot based on sub-dialogue models. *arXiv [preprint]. arXiv:2011.03259.*

Poppe, R., Truong, K. P., Reidsma, D., and Heylen, D. (2010). "Backchannel strategies for artificial listeners," in *Lecture Notes in Computer Science*, eds J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova (Berlin: Springer), 146–158. doi: 10.1007/978-3-642-15892-6_16

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue (Ph.D. thesis).* Department of Computer Sciencce, University of London, England. Available Online at: http://www.eecs.qmul.ac.uk/mpurver/papers/purver04thesis.pdf

Rajan, S., Craig, S. D., Gholson, B., Person, N. K., and Graesser, A. C. (2001). AutoTutor: incorporating back-channel feedback and other human-like conversational behaviors into an intelligent tutoring system. *Int. J. Speech Technol.* 4, 117–126. doi: 10.1023/A:1017319110294

Rodríguez, K. J., and Schlangen, D. (2004). "Form, intonation and function of clarification requests in german task-oriented spoken dialogues," in *Proceedings of SemDial 2004* (Barcelona), 101–108.

Romero-Trillo, J. (2019). Prosodic pragmatics and feedback in intercultural communication. *J. Pragmat.* 151, 91–102. doi: 10.1016/j.pragma.2019.02.018

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

Singh, N., Lee, J. J., Grover, I., and Breazeal, C. (2018). "P2PSTORY: dataset of children as storytellers and listeners in peer-to-peer interactions," in *Conference on Human Factors in Computing Systems - Proceedings* (Montréal, QC), 1–11. doi: 10.1145/3173574.3174008

Skantze, G. (2017). "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Proceedings of SIGdial* (Saarbrücken), 220–230. doi: 10.18653/v1/W17-5527

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: a review. *Comput. Speech Lang.* 67:101178. 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

Soldner, F., Pérez-Rosas, V., and Mihalcea, R. (2019). "Box of lies: multimodal deception detection in dialogues," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Minneapolis, MN), 1768–1777. doi: 10.18653/v1/N19-1175

Stivers, T. (2008). Stance, alignment, and affiliation during storytelling: when nodding is a token of affiliation. *Res. Lang. Soc. Interact.* 41, 31–57. doi: 10.1080/08351810701691123

Stocksmeier, T., Kopp, S., and Gibbon, D. (2007). "Synthesis of prosodic attitudinal variants in German backchannel JA, in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Antwerp), 409–412. doi: 10.21437/Interspeech.2007-232

Sun, M., Mou, Y., Xie, H., Xia, M., Wong, M., and Ma, X. (2019). Estimating emotional intensity from body poses for human-robot interaction. *arXiv [preprint]. arXiv:1904.09435.*

Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inform. Comput. Sci.* 43, 1947–1958. doi: 10.1021/ci034160g

Thepsoonthorn, C., Yokozuka, T., Miura, S., Ogawa, K., and Miyake, Y. (2016). Prior knowledge facilitates mutual gaze convergence and head nodding synchrony in face-to-face communication. *Sci. Rep*. 6, 1–14. doi: 10.1038/srep38261

Tolins, J., and Fox Tree, J. E. (2014). Addressee backchannels steer narrative development. *J. Pragmat*. 70, 152–164. doi: 10.1016/j.pragma.2014.06.006

Tozadore, D. C., and Romero, R. A. (2020). "Multimodal fuzzy assessment for robot behavioral adaptation in educational children-robot interaction," in *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Utrecht), 392–399. doi: 10.1145/3395035.3425201

Velentza, A. M., Heinke, D., and Wyatt, J. (2020). Museum robot guides or conventional audio guides? An experimental study. *Adv. Robot*. 34, 1571–1580. doi: 10.1080/01691864.2020.1854113

Verner, I. M., Polishuk, A., and Krayner, N. (2016). Science class with RoboThespian: using a robot teacher to make science fun and engage students. *IEEE Robot. Automat. Mag*. 23, 74–80. doi: 10.1109/MRA.2016.2515018

Visser, T., Traum, D., 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

Ward, N. G., and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmat*. 38, 1177–1207. doi: 10.1016/S0378-2166(99)00109-5

Werfel, J. (2014). "Embodied teachable agents: learning by teaching robots," in *New Research Frontiers for Intelligent Autonomous Systems*, (Padova, Italy), 1–8.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Comput. Linguist*. 35, 399–433. doi: 10.1162/coli.08-012-R1-06-90

Yamaoka, F., Kanda, T., Ishiguro, H., and Hagita, N. (2010). A model of proximity control for information-presenting robots. *IEEE Trans. Robot*. 26, 187–195. doi: 10.1109/TRO.2009.2035747

Yngve, V. H. (1970). "On getting a word in edgewise," in *CLS-70* (Chicago, IL: Chicago Linguistics Society), 567–578.

Yousuf, M. A., Kobayashi, Y., Kuno, Y., Yamazaki, A., and Yamazaki, K. (2012). "Developmen of a mobile museum guide robot that can configure spatial formation with visitors," in *International Conference on Intelligent Computing* (Huangshan: Springer), 423–432. doi: 10.1007/978-3-642-31588-6_55

Yu, C., and Tapus, A. (2019). "Interactive robot learning for multimodal emotion recognition," in *Lecture Notes in Computer Science*, eds A. M. Salichs, S. S. Ge, I. E. Barakova, J. J. Cabibihan, A. R. Wagner, A. Castro-Gonzalez, et al. (Madrid: Springer), 633–642. doi: 10.1007/978-3-030-35888-4_59

Zaletelj, J., and Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *Eurasip J. Image Video Process*. 2017, 1–12. doi: 10.1186/s13640-017-0228-8

Zhang, Y., Beskow, J., and Kjellström, H. (2017). "Look but don't stare: mutual gaze interaction in social robots," in *Lecture Notes in Computer Science*, 556–566. doi: 10.1007/978-3-319-70022-9_55

Zilka, L., and Jurcicek, F. (2016). "Incremental LSTM-based dialog state tracker," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings* (Scottsdale, AZ), 757–762. doi: 10.1109/ASRU.2015.7404864

**Conflict of Interest:** AA is a Ph.D. student at KTH's division of Speech, Music and Hearing. GS is a Professor in Speech Technology at KTH and co-founder of Furhat Robotics.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.