COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13392

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13392

Conversational Eyebrow Frowns Facilitate Question Identification: An Online Study Using Virtual Avatars

Naomi Nota,[a,b] James P. Trujillo,[a,b] Judith Holler[a,b]

[a] *Donders Institute for Brain, Cognition, and Behaviour, Nijmegen*
[b] *Max Planck Institute for Psycholinguistics, Nijmegen*

Received 7 September 2022; received in revised form 11 October 2023; accepted 27 November 2023

Abstract

Conversation is a time-pressured environment. Recognizing a social action (the ‘‘speech act,’’ such as a question requesting information) early is crucial in conversation to quickly understand the intended message and plan a timely response. Fast turns between interlocutors are especially relevant for responses to questions since a long gap may be meaningful by itself. Human language is multimodal, involving speech as well as visual signals from the body, including the face. But little is known about how conversational facial signals contribute to the communication of social actions. Some of the most prominent facial signals in conversation are eyebrow movements. Previous studies found links between eyebrow movements and questions, suggesting that these facial signals could contribute to the rapid recognition of questions. Therefore, we aimed to investigate whether early eyebrow movements (eyebrow frown or raise vs. no eyebrow movement) facilitate question identification. Participants were instructed to view videos of avatars where the presence of eyebrow movements accompanying questions was manipulated. Their task was to indicate whether the utterance was a question or a statement as accurately and quickly as possible. Data were collected using the online testing platform Gorilla. Results showed higher accuracies and faster response times for questions with eyebrow frowns, suggesting a facilitative role of eyebrow frowns for question identification. This means that facial signals

Correspondence should be sent to Naomi Nota, Donders Institute for Brain, Cognition, and Behaviour, Thomas von Aquinostraat 4, 6525 GD Nijmegen, The Netherlands. E-mail: Naomi.Nota@donders.ru.nl

can critically contribute to the communication of social actions in conversation by signaling social action-specific visual information and providing visual cues to speakers' intentions.

Keywords: Facial signals; Eyebrow movements; Social actions; Questions; Conversation; Turn-taking

1. Introduction

Conversation consists of rapid exchanges between interlocutors, involving minimal gaps and overlaps (Levinson & Torreira, 2015; Roberts, Torreira, & Levinson, 2015; Sacks, Schegloff, & Jefferson, 1974; Stivers et al., 2009). In order for this to be possible, it is important to recognize the speaker's social action (Levinson, 2013; a notion related to ‘‘speech act;’’ Austin, 1962; Searle, 1969) early to quickly understand the intended message (Gisladottir, Chwila, Schriefers, & Levinson, 2012, 2015, 2018; Holler & Levinson, 2019; Levinson, 2013). One of the most common and fundamental social actions in conversation is asking questions. Fast social action recognition is especially crucial for questions, since gaps longer than the average can indicate dispreferred responses (Schegloff, 2007), such as the declination of an offer, and are thus pragmatically marked (Bögels, Kendrick, & Levinson, 2015, 2020; Kendrick & Torreira, 2015).

It is now well established that human language is multimodal, consisting of bodily visual signals along with speech (e.g., Bavelas & Chovil, 2000; Holler & Levinson, 2019; Kendon, 2004; Levinson & Holler, 2014; McNeill, 1992, 2000; Perniss, 2018). Although the face is an important source of visual signaling in social interaction, facial signals have been primarily studied in the context of emotion. Only a few studies have focused on the role of facial signals in conversation with regard to some social actions. These studies suggest that beyond expressing emotion, facial signals may signal specific social actions in conversation (e.g., Bavelas, Gerwing, & Healing, 2014; Bavelas & Chovil, 2018; Nota, Trujillo, & Holler, 2021, 2022). An interesting avenue is to look at questions in particular, due to their fundamental role in conversation, and to investigate how facial signals may contribute to question identification. Therefore, in this study, we aim to investigate the contribution of eyebrow movements in questions, since they are prevalent facial signals in conversation (Nota et al., 2021, 2022).

1.1. Marking questions multimodally

Languages use different strategies to mark an utterance as a question in the spoken modality. These strategies are different between linguistic systems, and include variations in lexical items or morphemes, syntactic structure, or prosody. In many languages, questions are a marked sentence type. They are indicated by a specific word order, such as an inversion of the subject and verb by placing the verb at the beginning of the sentence (e.g., Dutch *Heeft hij het boek gelezen?* ‘‘Has he read the book?’’), specific question particles (e.g., French *est-ce-que*, English wh-words), or a combination of these strategies. Furthermore, common prosodic strategies to mark questions like high pitch and rising tune (e.g., Englert, 2010; Rossano, 2010) have been directly connected to questions (Hellbernd & Sammler, 2016; Sicoli, Stivers, Enfield, & Levinson, 2015). Thus, there are both linguistic and prosodic ways to mark questions in the spoken modality.

Several studies showed that the visual modality may also play a role in signaling questions. For instance, research investigating speech production in various spoken and signed languages frequently observed an association between eyebrow movements and questions (Bavelas et al., 2014; Borràs-Comes, Kaland, Prieto, & Swerts, 2014; Chovil, 1991; Coerts, 1992; Domaneschi, Passarelli, & Chiorri, 2017; Ekman, 1979; Hömke, Holler, & Levinson, 2019, 2022; Nota et al., 2021, 2022; Torreira & Valtersson, 2015; Zeshan, 2004). Specifically, both eyebrow frowns and eyebrow raises were found to frequently co-occur with questions (Borràs-Comes et al., 2014; Chovil, 1991; Domaneschi et al., 2017; Ekman, 1979; Torreira & Valtersson, 2015). Furthermore, a recent study examining Dutch conversations showed that eyebrow frowns were among the most typical visual question markers compared to other facial signals (Nota et al., 2021), as well as strong markers of information requests compared to other social actions performed by questions (Nota et al., 2022). These previous studies show that eyebrow frowns and eyebrow raises may be particularly relevant signals to mark questionhood.

This is in line with experimental research investigating speech comprehension, where the availability of auditory cues (e.g., F0 contour, intensity, and pitch range) and visual signals like eyebrow frowns and raises were manipulated to study their individual contribution. Both auditory cues and visual signals were found to distinguish questions from statements (Borràs-Comes et al., 2014; Borràs-Comes & Prieto, 2011; Cruz, Swerts, & Frota, 2017; House, 2002; Miranda, Swerts, Moraes, & Rilliard, 2021; Srinivasan & Massaro, 2003; Torreira & Valtersson, 2015), showing that visual signals may contribute to signaling questionhood. Visual signals may also help to differentiate between different questions, since both auditory cues and visual signals distinguished polar (yes-no) questions that echoed (a part of) preceding speech from incredulity questions (comparison of Dutch and Catalan in Crespo Sendra, Kaland, Swerts, & Prieto, 2013).

In some comprehension studies, however, visual signals alone were found to be less reliable indicators of questions compared to auditory cues only (Cruz et al., 2017; House, 2002; Miranda et al., 2021; Srinivasan & Massaro, 2003), even when the signals were enhanced (Srinivasan & Massaro, 2003). Conversely, auditory cues *combined* with visual signals were better indicators of questions compared to auditory cues only (Borràs-Comes et al., 2014; Crespo Sendra et al., 2013). However, the effect of eyebrow movements accompanying questions may also vary with language. Languages differ in their question marking strategies, which might have influenced which strategies participants more dominantly relied on. It could be that question marking in the visual modality is used more when the verbal utterance is lacking a lexico-syntactic structure typically found in questions (e.g., subject–verb inversion, wh-fronting). This was proposed to be the case for English (Ekman, 1979), and was demonstrated in research comparing Dutch and Catalan, where Dutch participants relied more on intonational differences, whereas Catalan participants relied more on visual signals (Borràs-Comes et al., 2014; Crespo Sendra et al., 2013). In Dutch, polar questions are marked by subject–verb inversion, whereas Catalan does not use this specific morphosyntactic strategy, but instead relies on a particular prosodic contour consisting of a low pitch accent followed by a rising boundary tone (Borràs-Comes et al., 2014; Prieto et al., 2015). This is similar to the trade-off relationship observed between visual signals and auditory cues in Italian (Rossano, 2010). Rossano (2010) found more visual signals for polar questions and alternative questions compared to wh-questions, which was explained by the fact that the latter use more morphosyntactic cues than the other types of questions.

Irrespective of language differences, it could be that different eyebrow movement patterns exist for different question types. This is true for many sign languages, where eyebrow movements are grammatical markers of questions (Zeshan, 2004). For instance, in Dutch sign language (Nederlandse Gebarentaal; NGT), eyebrow raises typically mark polar questions, while eyebrow frowns mark content questions (Coerts, 1992). Similarly, in spoken Dutch, eyebrow raises mark open requests for repair (e.g., *What?; Huh?*) and restricted offers to request clarification (e.g., *Johnny Smith?*), while eyebrow frowns mark restricted requests for repair (e.g., *Who?; He did what?*; Hömke et al., 2019, 2022). Even within one language, question type may thus be an important factor to consider in analyses of the role of eyebrow movements when speakers produce questions and when recipients comprehend them.

In short, there seems to be evidence from a confluence of studies of spoken language and experiments testing comprehension that eyebrow movements may play a role in making questions recognizable. However, none of those previous studies have taken detailed features of naturalistic conversation into account. First, this concerns differences in lexical or syntactic structure of questions, which may influence the extent to which eyebrow movements fulfil a signaling function in the recognition of questionhood. Second, and perhaps more importantly, none of the studies to date have considered the effect of eyebrow movements on the time course of question identification. As laid out above, time is of the essence in conversational turn-taking, and identifying questions quickly is paramount, due to the pragmatic marking of slow responses to questions (Schegloff, 2007; Kendrick & Torreira, 2015). Especially visual signals that start before or at the onset of the verbal utterance may be particularly helpful for fast social action recognition in a turn-taking context (Holler & Levinson, 2019). Questions with hand gestures occurring before or with questions appear to be associated with faster responses, which lead to speculations about mechanisms resulting in multimodal facilitation (ter Bekke et al., 2020; Holler, Kendrick, & Levinson, 2018). Thus, a similar effect may be observable for facial signals like eyebrow movements, especially since these often occur prior to or around the onset of the utterance (Kaukomaa, Peräkylä, & Ruusuvuori, 2014; Nota et al., 2021, 2022). Their early timing could help the addressee understand the speaker's intended message more quickly by speeding up recognition of the social action. Especially interesting, therefore, are two things relating to timing: (a) measuring the speed of question identification in the presence and absence of eyebrow movements, as well as how different forms of eyebrow movements may affect response speed, and (b) measuring the effect of how early an eyebrow movement occurs in relation to verbal question onset on the speed of identifying questions. The present study will address these outstanding issues.

1.2. Current study

The current study addresses the following research question: What is the influence of eyebrow movements on accuracy and on speed of question identification? To answer this question, we created avatars and manipulated the presence of eyebrow movements accompanying questions. We looked not only at accuracy scores but, crucially, also at response times (RTs) in a behavioral two alternative forced choice experiment where participants viewed videos of avatars and indicated whether the utterance was a question or statement.

An important feature of the present study is that the spoken and visual materials were based on naturally produced conversations from a face-to-face Dutch conversation corpus. All utterances were either information request questions, as they were the most frequently occurring questions performed in the corpus (Nota et al., 2022), or statements, since statements are generally unmarked sentence types, and are often used to convey information. Information request questions served as the critical utterances, while statements were used as fillers. We selected early eyebrow movements (frowns or raises), since they most frequently occurred in the corpus (Nota et al., 2021, 2022). These eyebrow movements had different onset times occurring before or at the start of the verbal utterance, and different durations. We also controlled for linguistic structure of the question and turn duration. Furthermore, different psycho-social questionnaires were collected to provide a more comprehensive characterization of the sample.

We hypothesized that there would be a main effect of eyebrow movement presence, with higher accuracies and faster RTs for questions with eyebrow movements compared to questions without eyebrow movements. We expected eyebrow frowns to have a facilitative effect on question identification, in line with the observed association between eyebrow frowns and questions in the Dutch conversation corpus (Nota et al., 2021, 2022). We additionally expected that the facilitative effect could also occur for eyebrow raises, due to the observed association between eyebrow raises and questions in the previous literature, but our expectation here was not as strong due to the common occurrence of eyebrow raises in other contexts (such as for prosodic emphasis; Swerts & Krahmer, 2008) that may make them weaker signals for any specific social action.

This study extends prior research showing the importance of eyebrow movements for the identification of questions in Dutch. Specifically, we captured the effect of visual signaling not only on accuracy but also on the speed of question identification, by using eyebrow movements of timings that are representative of spontaneous multimodal behavior. This study marks a critical departure from previous studies traditionally using confederates or actors since the present study used virtual avatars to collect controlled data from a large online sample. Thus, this study combines experimental control with capturing naturalistic multimodal behavior to understand the role of facial signals in social signaling and conversational face-to-face interaction. Our findings provide novel insights into whether facial signals critically contribute to the communication of fundamental social actions in conversation by signaling social action-specific visual information, and providing visual cues to speakers' intentions.

2. Methods

The preregistration of this study is available on the As Predicted website https://aspredicted.org/1T9_H8R. A comprehensive preregistration, power analysis, analysis scripts with session information, depersonalized data, results, and supplementary materials are available on the Open Science Framework project website https://osf.io/xjpaq/.

2.1. Participants

Ninety native speakers of Dutch between 18 and 45 years old (mean age: 22 ± 5 years) with no motoric, hearing, or language problems and normal or corrected-to-normal vision were recruited through the subject database of the Radboud University in Nijmegen. A number of participants ($n = 10$) were excluded from the final sample manually because they did not complete the experiment or did not meet the study requirements. This resulted in a final sample of 80 participants (71 females, 9 males). This chosen sample size was based on a power analysis performed on a pilot study of 10 participants (mean age: 22 ± 4 years, seven females, three males) that indicated that 80 participants would provide sufficient statistical power ($\beta > .8$) for a small effect size ($d = .2$). Participants gave written informed consent prior to the study and were paid 7.50 euros at the end of the experiment. The study was approved by the Ethics Committee of the Social Sciences department in the Netherlands.[1]

To characterize our sample, we assessed participants' Empathy Quotient (EQ) score with the Dutch version of the EQ questionnaire (Baron-Cohen & Wheelwright, 2004; Groen, Fuermaier, Heijer, Tucha, & Althaus, 2016), in which participants need to self-report whether they agree with 60 statements relating to empathy on a four-point scale ranging from (1) strongly disagree to (4) strongly agree. The Actions and Feelings Questionnaire (AFQ) score was measured with the Dutch version of the AFQ questionnaire (van der Meer, Sheftel-Simanova, Kan, & Trujillo, 2021; Williams & Cameron, 2017), in which participants need to self-report whether they agree with 18 statements capturing self-awareness of actions related to feelings, on a four-point scale ranging from (1) strongly disagree to (4) strongly agree. Furthermore, avatar evaluation was estimated by asking participants whether they agree with three statements designed to assess the participants' perception of the avatars' (a) humanness, (b) ease of understanding, and (c) likeability. These statements were: (a) *Ik vond de avatars menselijk overkomen* (''These avatars appeared human''), (b) *Ik kon de avatars makkelijk begrijpen* (''I think these avatars were easy to understand''), and (c) *Ik vond deze avatars sympathiek overkomen* (''These avatars appeared nice''). Participants were instructed to self-report whether they agree with the statements on a six-point scale ranging from (1) strongly disagree to (6) strongly agree (based on Heyselaar, Hagoort, & Segaert, 2017a; Hömke et al., 2019; Weatherholtz, Campbell-Kibler, & Jaeger, 2014). An overview of participants' EQ scores, AFQ scores, avatar evaluations, and an exploratory analysis on their effects on accuracy and RT are provided in the Appendix.

2.2. Stimuli

The avatar stimuli were based on (frontal) audio and video recordings from a corpus of naturalistic face-to-face Dutch conversations (CoAct corpus). The corpus consisted of pairs of acquaintances holding a casual conversation for 1 h while being recorded. Questions and responses, social actions of questions (e.g., information requests, which ask for new information of factual or specific nature), question types (polar or content, as well as types of polar questions), and facial signals (like eyebrow frowns and raises) of the speakers in the corpus were manually transcribed by Dutch speakers using ELAN (5.5; Sloetjes & Wittenburg, 2008; for more details on the corpus conventions, see Nota et al., 2021, 2022; Trujillo & Holler, 2021). The transcription of questions and responses largely followed the coding scheme of Stivers and Enfield (2010), with additional rules to account for the complexity of the corpus data. A holistic coding approach was adopted, in which the conversational context, communicative intention, and interactional response were used to segment and identify questions and responses in the corpus. We then used these utterances from the corpus as our stimuli. For questions and responses, interrater reliability between coders was assessed on 12% of the data with raw agreement, a modified Cohen's kappa (Cohen, 1960; Landis & Koch, 1977) using EasyDiag (Holle & Rein, 2015), and a standard overlap criterion of 60%. This resulted in a raw agreement of 75% and $k = .74$ for questions, as well as 73% and $k = .73$ for responses, indicating substantial agreement. A subset of 2082 questions were coded for their social action category and question type. Interrater reliability was measured on 686 additionally coded social action categories of the question annotations following the same procedure as for questions and responses. This resulted in a raw agreement of 76% and $k = .70$ for social actions of questions, indicating substantial agreement. For question types and polar question types, raw agreement was 98% and $k = .97$, indicating almost perfect agreement. Facial signals were annotated when they involved movements that carried some form of communicating meaning related to the questions and responses. For facial signals, interrater reliability was calculated by randomly selecting one question and one response in one of the three recording parts for each participant in all dyads ($n = 136$, roughly equivalent to 1% of the question and response data in the entire corpus) resulting in the additional annotation of 223 facial signals for reliability, and allowing a pairwise comparison between coders. We chose this pairwise analysis due to the unequal amount of data that was transcribed between coders. The paired comparisons of the facial signals showed an average raw agreement of 76%, and an average $k = .96$, indicating almost perfect agreement (Landis & Koch, 1977).

For this investigation, we used a total of 80 questions consisting of 40 polar questions (18 interrogatives, 10 declaratives, 11 tag-questions, and 1 non-clausal question) and 40 content questions (all wh-questions). All questions were information requests, which could be, for example, requests of factual nature ("What kind are they?"), non-factual nature ("What do you think?"), requests for confirmation of information ("You saved the date in your calendar right?"), or requests for elaboration ("And then?"). Forty questions had eyebrow frowns (20 polar, 20 content), and 40 questions (20 polar, 20 content) had eyebrow raises (see Table 1 for more information about the questions subdivision). Questions without eyebrow movements were rendered by stripping away the original eyebrow movements from the questions with eyebrow frowns and raises. Thus, the original eyebrow movements were recreated based on the corpus for one condition and left out in the other condition. Since the stimuli consisted of extracted utterances from the corpus, the stimuli were segmented and defined while taking into account the conversational context they appeared in.

We additionally used 80 statements as fillers to distract participants from the main research question. Twenty statements had eyebrow frowns, and 20 statements had eyebrow raises. Forty statements did not have eyebrow movements. These statements were selected from the

Table 1
Overview of the questions' social action subdivision

	Eyebrow movements			
Questions (all information requests)	40 Frowns		40 Raises	
	20 Polar	20 Content	20 Polar	20 Content
	9 Interrogatives 7 Declaratives 4 Tag-	20 Wh-	9 Interrogatives 3 Declaratives 7 Tag- 1 Non-clausal	20 Wh-

same speakers as the questions were selected from, in order to obtain an identical amount of questions and statements from each speaker. Furthermore, the statements with an eyebrow movement (frown or raise) had the same type of eyebrow movement as the questions from the same speakers (e.g., if a question with a frown was selected from speaker 1, a statement with a frown was selected from speaker 1 as well). There were another 20 additional fillers (10 questions and 10 statements) accompanied by facial signals of the eyes other than eyebrow movements (five eye widenings, five squints), to vary the visual signals and distract from the experimental aim.

Since most facial signals frequently occurred before or at the start of the verbal utterance in the corpus (Nota et al., 2021, 2022), we only selected information request questions and statements with eyebrow frowns, eyebrow raises, eye widenings, or squints that occurred before or at the onset of the utterance (mean onset difference between visual signals and verbal utterances: 536 ms). Within this time window, the facial signal movement onsets were variable. This resulted in the inclusion of recordings of 53 individual speakers from the corpus (mean age: 22 ± 7 years, 41 females, 12 males) for the avatar animations.

2.3. Avatars

The corresponding video clips of each stimulus were retrieved from the corpus by converting the mp4 recordings to mp4 iframes, cutting them, and exporting the speech (wav) using ffmpeg (4.3.1; Tomar, 2006) with custom scripts in Bash (3.2.51-1) and Ruby (2.3.0). The avatars were created in Blender (2.83.12; Community, 2018) using the open-source plug-in MB Lab (1.7.6, Bastioni, 2021) for the parametric 3D modeling of humanoid characters. The appearance of the avatars was matched to the original corpus speakers in terms of skin, hair, and clothing. Face detection was based on action units, gaze, and pose of the corpus video clips retrieved by OpenFace (2.2.0; Baltrusaitis, Zadeh, Lim, & Morency, 2018). This was imported in Blender with the open-source add-on FACSvatar (0.3.4-alpha; van der Struijk, Huang, Mirzaei, & Nishida, 2018) using the FACSvatar-Blender plug-in (0.4.0). FACSvatar is a modular framework for real-time facial action coding system (FACS)-based facial animation (Ekman & Friesen, 1978). In addition to the facial signals in the stimuli, the avatars showed direct gaze and blinks, which were added manually throughout all utterances. With the exception of gaze, the timing of all facial signals was extracted from the facial signal transcriptions of the corpus (for more details on the corpus conventions, see Nota et al., 2021). The intensities of these facial signals were fixed, and determined by using the standard intensities generated via FACSvatar (0.3.4-alpha; van der Struijk et al., 2018). Because the facial structure for the male and female avatars was different, the intensity values had to be slightly adjusted to lead to comparable appearances and look visually proportionate between the male and female versions (see Supporting Information for the specific intensities, and a histogram showing the mean structural similarity score between eyebrow frown and eyebrow raise movements, demonstrating very little difference between the two signal types).

Finally, the avatars' lip movements were synchronized with the speech files using a custom script in Python (3.7) with phoneme annotations, and manually edited. The precise beginnings and endings of the phonemes were segmented in Praat (5.1, Broersma & Weenink, 2001) using the automatic segmentation tool WebMAUS Basic (3.3; Kisler, Reichel, & Schiel, 2017) and manual transcription. These transcriptions were imported to their corresponding file in ELAN (5.5; Sloetjes & Wittenburg, 2008), and exported as text files.

2.4. *Design*

The stimuli were pseudorandomized and balanced across two lists. The lists allowed us to present questions in two conditions: (a) with eyebrow movement (frown, raise) and (b) without eyebrow movement (frown or raise removed). This was counterbalanced in the two lists, so that each participant saw each question only in one condition. Statements were always the same set in the two lists and were not manipulated in the same way as questions were (i.e., questions without eyebrow movements were made by removing eyebrow movements from questions with eyebrow frowns and raises), since there were not enough statements with eyebrow movements in the original corpus to create different sets. Thus, all participants saw each statement with the same eyebrow movement (or lack thereof) in the two lists. Like statements, all other fillers (questions and statements with eye widenings or squints) were always the same in the two lists. The two lists had a total of 180 trials. The experiment consisted of five blocks per list (with 16 questions, 16 statements, 4 fillers per block, and with a balanced number of utterances with and without visual signal per block), allowing participants to rest between blocks. The order of blocks was randomized, as well as the items within each block. Table 2 provides an overview of the experimental design.

Each video clip began with a still frame of a neutral face (without any visual signal) lasting 400 ms, by adding 10 single frames (25 fps) to the first video frame with ffmpeg (4.3.1; Tomar, 2006). If the first video frame was not a neutral face, we looked for another frame with a neutral face manually. This avoided the impression of an abrupt start. After that, all visual signals started before or at the onset of the verbal utterance (determined by the timing in the corpus), and had a gradual fade in and fade out that was largely based on the original fade lengths of those signals. Facial signal fades were coded in ELAN (5.5; Sloetjes & Wittenburg, 2008) from the first evidence of movement until the movement peak, or from the movement peak until the last evidence of movement. Fades under two frames were changed to 80 ms,

Table 2
Overview of the experimental design

	Eyebrow movements		No eyebrow movements	
	Frowns	Raises	Frowns removed	Raises removed
Questions	20 (10 polar, 10 content)	20 (10 polar, 10 content)	20 (10 polar, 10 content)	20 (10 polar, 10 content)
	Frowns	Raises	No eyebrow movements	
Statements	20	20	20	20

to make the gradual build-up of the visual signals look more natural. After the 10 frames of neutral face, videos continued at the first frame of eyebrow movement. Specifically, for questions with a removed eyebrow movement, video clips continued at the original eyebrow movement onset, to keep durations consistent across conditions. For statements without any eyebrow movement, video clips continued at the eyebrow movement onset from a randomly paired statement with an eyebrow movement, since there was no original eyebrow movement onset. This meant the length of time the avatar was shown before speech onset was the same as for the randomly paired statement. The verbal utterance started at the original timing. To avoid distraction from potential background noise, we muted the sound from the beginning of the video clip until the start of the verbal utterance using ffmpeg (4.3.1; Tomar, 2006). Additionally, the overall sound intensity was normalized to 60 dB sound pressure level and converted to mono in Praat (version 6.1.40; Broersma & Weenink, 2021), to prevent loudness differences between items. See Fig. 1 for an example of a video clip in two conditions.

2.5. Procedure

Participants were asked to install the most recent version of Google Chrome and use Windows, before launching the experiment on a laptop or desktop using the online experiment platform Gorilla (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). These criteria were used because this combination of browser and operating system shows the smallest and most consistent visual delay across different platforms (Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2020). Furthermore, we instructed participants to use wired earphones or headphones, to avoid potential sound delays from Bluetooth devices. To make sure these instructions were followed, we limited participants' browser and device via Gorilla's graphical user interface in order to exclude participants who did not use Google Chrome and a computer. Moreover, we rejected participants who did not meet the study requirements through a short questionnaire that asked about their browser, device, and earphone or headphone type at the start of the experiment. Participants who met the study requirements continued with a general demographics and language background questionnaire.

Before starting with the experimental task, we asked participants to sit in a quiet room behind a desk or table, switch off their phones and other electronics (e.g., television), turn off their notifications (e.g., email), and put on headphones or earphones for the duration of

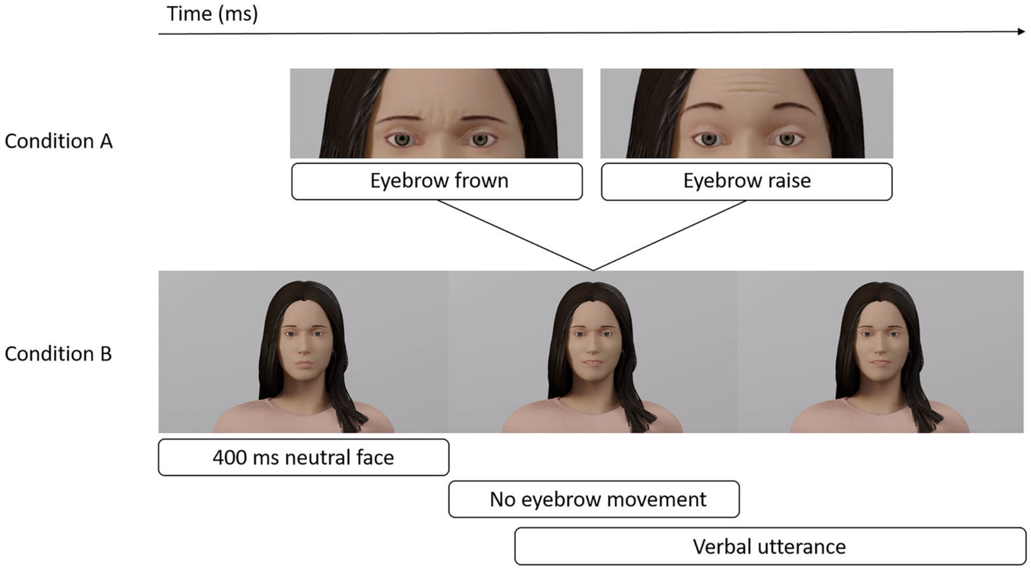

Fig 1. Example of a video clip. The video clip consists of an utterance with eyebrow movement (condition A) and without eyebrow movement (condition B). The mean duration of the video clip (in ms) is the same across conditions. Example videos of the different conditions are available on the Open Science Framework project website https://osf.io/xjpaq/.

the experiment. Then, a sound was played to allow them to adjust their volume, and to make sure that autoplay was enabled on their browsers. This was an adaptation of the web-based headphone screening test (Milne et al., 2020). The experimental task opened in full-screen automatically to make sure that the video clips (1280 × 720) scaled to the maximum space that was available on the participants' monitor size.

Participants were explicitly instructed to indicate whether the utterance was a question (*vraag*) or a statement (*stelling*) by using response keys "X" and "M" as quickly and accurately as possible. These keys were selected since their placement is the same between different keyboards. The 180 experimental trials consisted of video clips of the avatars and were preceded by 16 practice items (eight questions and eight statements, each with one eyebrow frown, one eyebrow raise, two without eyebrow movements, two squints, and two eye widenings). Before each trial, a fixation cross appeared at a fixed position in the middle of the screen for 500 ms, after which the video clip was played. As soon as there was a button-press, a blank screen was presented for 1000 ms, and then the next trial began. If there was no button-press, the video clip played in its entirety. The last video frame was then shown until there was a button-press before moving to the blank screen. RT was measured from the beginning of the video until the button-press. There were four self-paced breaks between blocks. Participants' accuracy and RTs were recorded. Fig. 2 provides an example of an experimental trial.

Upon completion of the experiment, participants were instructed to fill in four questionnaires. The order of the first two questionnaires was randomized. These were the EQ (Baron-Cohen & Wheelwright, 2004) and the AFQ (van der Meer et al., 2021; Williams & Cameron,

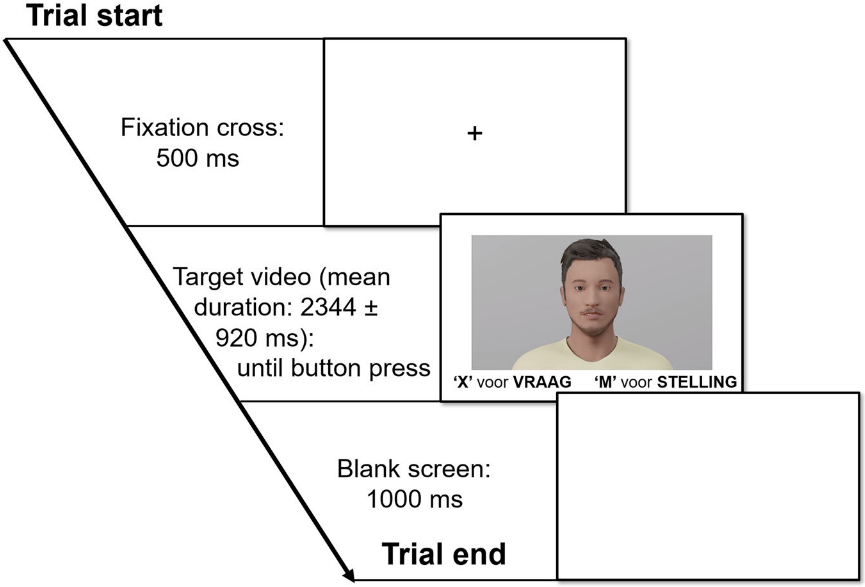
Trial start
Fixation cross: 500 ms
+
Target video (mean duration: 2344 ± 920 ms): until button press
'X' voor VRAAG 'M' voor STELLING
Blank screen: 1000 ms
Trial end

Fig 2. Example of an experimental trial.

2017). The third questionnaire was the avatar evaluation, and the fourth questionnaire consisted of questions assessing explicit awareness of the experimental aim (see above for details on the first three questionnaires). The entire experiment lasted approximately 45 min.

2.6. *Analysis*

To answer our research question, we analyzed participants' accuracy scores and RTs. First, we looked whether participants were aware of the experimental aim. While several participants recognized that the study was about visual signals, none were aware of the specific research focus. Then, the data were checked for outliers, which we considered to be RTs more than 2.5 *SD* from the mean participant RT. This resulted in the removal of 189 trials (2.95% of total trials). Inaccurate responses were excluded from the RT dataset.

We used generalized linear mixed-effect models (GLMMs) in *R* (4.1.2; R Core Team, 2021) with *RStudio* (2021.09.2-382; RStudio Team, 2022). GLMMs allow for the inclusion of additional predictors, and provide a solution for fitting the distributions by satisfying normality assumptions without the need for transformation (Lo & Andrews, 2015). The models were run with the *glmer* function as implemented in the *lme4* package for *R* (1.1-28; Bates, Mächler, Bolker, & Walker, 2015). Following the recommendations of Meteyard and Davies (2020), we selected the fixed and random parameters of our models on the basis of our research questions and experimental design. This has the added benefit of reducing the chance of overfitting.

For each of the two dependent variables (accuracy, RT), the fixed effects were eyebrow movement (with frown, with raise, without eyebrow movement), the difference between eyebrow movement onset and utterance onset (scaled), eyebrow movement duration (scaled), question type (declarative, interrogative, non-clausal, tag-, wh-), and utterance duration (scaled). The reference levels were absence of eyebrow movement and wh-question type. We included random intercepts by participant and item, and did not add random slopes since this led to convergence issues in the power analysis based on pilot data, and would unnecessarily add complexity to the models (Meteyard & Davies, 2020).

We ran log-likelihood ratio tests (ANOVA function) for accuracy and RT to test for the presence of main effects for all fixed parameters. To explore whether there was a differential use of eyebrow movements depending on the lexico-syntactic structure of questions, we ran log-likelihood ratio tests for accuracy and RT that tested for the presence of an interaction between eyebrow movement and question type. In all models, we applied Helmert contrasts for the eyebrow movement variable, so that the first contrast compared absence of eyebrow movement with presence of eyebrow frown, and the second contrast compared absence of eyebrow movement with presence of eyebrow raise. Our experiment design and mixed modeling approach ensured that when we compared absence of eyebrow movement against eyebrow frown or raise, we specifically compared the same utterances accompanied by no eyebrow frowns versus eyebrow frowns, and the same utterances accompanied by no eyebrow raises versus eyebrow raises.

We additionally performed a post-hoc analysis for eyebrow movement, question type, and the interaction between eyebrow movement and question type using the Tukey method with *emmeans* (1.7.2; Lenth, 2019).

3. Results

3.1. Accuracy

3.1.1. Effect of eyebrow movements

For accuracy, there was a significant main effect of eyebrow movement ($\chi^2(2) = 22.67, p < .001$). There were no significant main effects of difference in onsets and of eyebrow duration for accuracy.

The post-hoc analysis revealed a significant difference between eyebrow frowns and no eyebrow movement ($\beta = -.61, SE = .13, z = -4.72, p < .001$), but no significant difference between eyebrow raises and no eyebrow movement. This shows that only the presence of eyebrow frowns resulted in higher accuracies compared to the absence of eyebrow movement, but not the presence of eyebrow raises. Fig. 3 shows an overview of the accuracy results (an overview of the accuracy results per participant is available on the Open Science Framework project website https://osf.io/xjpaq/).

3.1.2. Effect of question type and utterance duration

There was a significant effect of question type ($\chi^2(4) = 61.87, p < .001$), but no significant effect of the interaction between eyebrow movement and question type. Moreover, there was

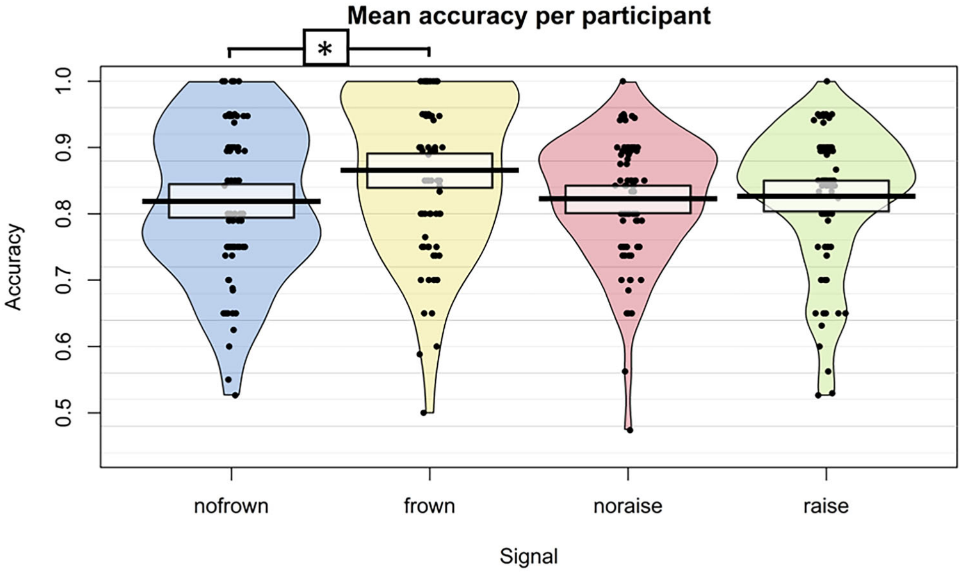

Fig 3. Overview of the accuracy results. For better visualization, absence of eyebrow movement is split into absence of eyebrow frown and absence of eyebrow raise. Pirate plots depict the distribution of the accuracies for absence of eyebrow frown, eyebrow frown, absence of eyebrow raise, and eyebrow raise. Individual dots represent overall mean accuracy for individual participants (raw data). Bars indicate means, beans (the oval shapes around the dots) indicate smoothed density, and bands (dark-colored lines at the top of the bars) indicate the 95% Bayesian highest density interval (HDI). Asterisks indicate a significant difference between conditions. The model equation was: Accuracy ~ Eyebrow movement (eyebrow frown, eyebrow raise, no eyebrow movement) + Difference onsets + Eyebrow movement duration + Question type (declarative, interrogative, non-clausal, tag-, wh-) + Utterance duration + (1 | Participant) + (1 | Item).

a significant effect of utterance duration ($\chi^2(1) = 10.33$, $p = .001$), showing that a longer utterance resulted in higher accuracies.

The post-hoc analysis on the effect of question type on accuracies revealed significantly higher accuracies for interrogative questions compared to declarative questions ($\beta = -3.34$, $SE = .46$, $z = -7.22$, $p < .001$) and tag-questions ($\beta = 2.36$, $SE = .46$, $z = 5.17$, $p < .001$), as well as significantly higher accuracies for wh-questions compared to declarative questions ($\beta = -3.37$, $SE = .42$, $z = -8.05$, $p < .001$) and tag-questions ($\beta = -2.39$, $SE = .41$, $z = -5.79$, $p < .001$).

3.2. Response time

3.2.1. Effect of eyebrow movements

For RT, there was a significant main effect of eyebrow movement ($\chi^2(6) = 23.21$, $p < .001$). Furthermore, there was a significant main effect of difference in onsets ($\chi^2(1) = 75.95$, $p < .001$), showing that a larger difference in onsets (i.e., earlier eyebrow movements) resulted in faster RTs. Moreover, there was a significant main effect of eyebrow duration ($\chi^2(1) = 5.76$, $p = .016$), showing that a longer eyebrow duration resulted in faster RTs.

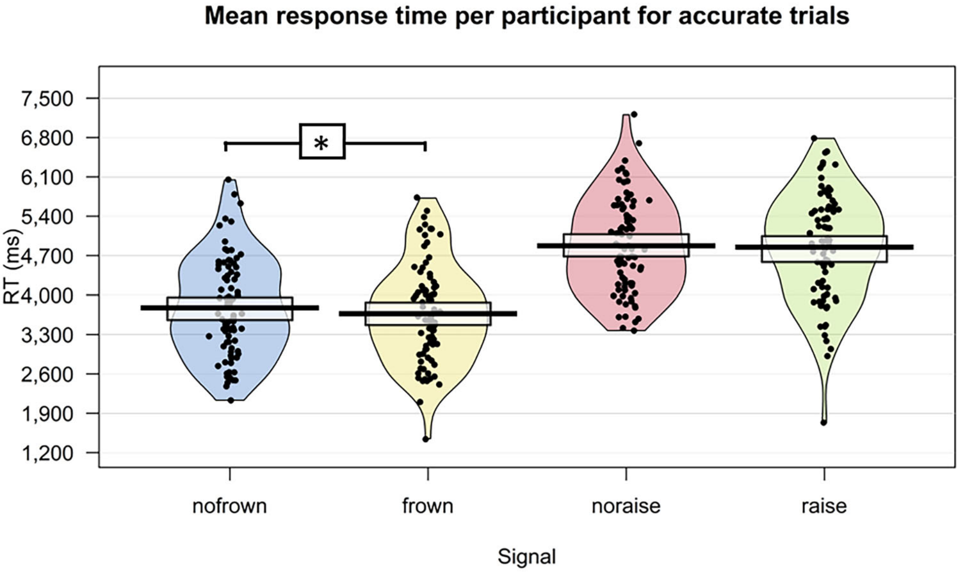

Fig 4. Overview of the RT results. For better visualization, absence of eyebrow movement is split into absence of eyebrow frown and absence of eyebrow raise. Pirate plots depict the distribution of the accuracies for absence of eyebrow frown, eyebrow frown, absence of eyebrow raise, and eyebrow raise. Individual dots represent overall mean RT for individual participants (raw data). Bars indicate means, beans (the oval shapes around the dots) indicate smoothed density, and bands (dark-colored lines at the top of the bars) indicate the 95% Bayesian Highest Density Interval (HDI). Asterisks indicate a significant difference between conditions. The model equation was: RT $\sim$ Eyebrow movement (eyebrow frown, eyebrow raise, no eyebrow movement) + Difference onsets + Eyebrow movement duration + Question type (declarative, interrogative, non-clausal, tag-, wh-) + Utterance duration + (1 | Participant) + (1 | Item). RT was measured from the beginning of the video until the button press.

The post-hoc analysis for eyebrow movement revealed a significant difference between eyebrow frowns and no eyebrow movements ($\beta = .04$, $SE = .01$, $z = 2.70$, $p = .019$), but no significant difference between eyebrow raises and no eyebrow movement. This shows that only the presence of eyebrow frowns resulted in faster RTs compared to the absence of eyebrow movement, but not the presence of eyebrow raises. Fig. 4 shows an overview of the RT results (an overview of the RT results per participant is available on the Open Science Framework project website https://osf.io/xjpaq/).

3.2.2. *Effect of question type and utterance duration*

There was a significant effect of question type ($\chi^2(4) = 15.34$, $p = .004$), but no significant effect of the interaction between eyebrow movement and question type. Moreover, there was a significant effect of utterance duration ($\chi^2(1) = 28.05$, $p < .001$), showing that a longer utterance resulted in longer RTs.

The post-hoc analysis on the effect of question types on RTs revealed significantly faster RTs for interrogative questions compared to declarative questions ($\beta = .25$, $SE = .05$, $z = 5.24$, $p < .001$) and tag-questions ($\beta = -.26$, $SE = .04$, $z =$ -5.96, $p < .001$), as well as significantly faster RTs for wh-questions compared to declarative questions ($\beta = .17$, $SE = .04$, $z = 3.87$, $p = .001$) and tag-questions ($\beta = .18$, $SE = .04$, $z = 4.58$, $p < .001$).

4. Discussion

In the current study, we investigated the influence of eyebrow movements on accuracy and on speed of question identification. We created avatars and manipulated the presence of eyebrow movements accompanying questions, and looked at accuracy scores and RTs from a behavioral two-alternative forced choice experiment where participants viewed videos of the avatars and indicated whether the utterance was a question or statement.

Results showed significantly higher accuracies for questions with eyebrow frowns compared to questions without eyebrow movements. However, there was no significant difference in accuracies between questions with eyebrow raises compared to questions without eyebrow movements. Questions with eyebrow frowns also had significantly faster RTs compared to questions without eyebrow movements. No significant differences in RTs were observed between questions with eyebrow raises and questions without eyebrow movements. A larger difference between eyebrow movement onset and utterance onset resulted in significantly faster RTs, and a longer eyebrow duration resulted in faster RTs. Moreover, specific question types had an effect on accuracy and RT. Both interrogative and wh-questions resulted in higher accuracies and faster RTs than declarative and tag-questions. Finally, longer utterances resulted in higher accuracies but longer RTs.

The finding that questions with eyebrow frowns resulted in higher accuracies and faster RTs compared to questions without eyebrow movements is in line with previous studies showing an association between eyebrow movements and questions (Bavelas et al., 2014; Borràs-Comes et al., 2014; Chovil, 1991; Coerts, 1992; Domaneschi et al., 2017; Ekman, 1979; Hömke et al., 2019, 2022; Nota et al., 2021, 2022; Torreira & Valtersson, 2015; Zeshan, 2004), and studies showing that eyebrow movements help to identify questions (Borràs-Comes et al., 2014; Borràs-Comes & Prieto, 2011; Cruz et al., 2017; House, 2002; Miranda et al., 2021; Srinivasan & Massaro, 2003; Torreira & Valtersson, 2015). Thus, this study replicates previous findings, and builds on prior research by showing a facilitative role of facial signals for the accurate detection of questions when using eyebrow frowns with timings that are mapped onto spontaneous multimodal behavior.

The faster RTs to questions with eyebrow frowns further support the idea that early visual signaling cues the social action, and may, in turn, facilitate early recognition of the speaker's intention (Gisladottir et al., 2012, 2015, 2018; Holler & Levinson, 2019; Levinson, 2013). This is similar to the multimodal facilitation observed for questions with early manual and head gestures (ter Bekke et al., 2020; Holler et al., 2018). The early visual signaling facilitation effect is further demonstrated by the faster RTs observed for larger differences in eyebrow onsets and verbal utterance onsets, showing that participants responded faster to eyebrow movements that occurred earlier before speech.

Moreover, longer eyebrow movements resulted in faster RTs, suggesting that especially longer facial signals are clearer cues for indicating questionhood, potentially due to the fact that they are easier to perceive than shorter ones. Thus, it may be that long facial signals that foreshadow the verbal utterance are especially helpful in signaling social action-specific visual information. This is comparable to previous research where eyebrow frowns were observed to foreshadow an utterance (Kaukomaa et al., 2014). Early facial signals may therefore facilitate prediction of the speakers' intentions and allow for the next speaker to produce a timely response. Fast responding is crucial in a turn-taking context, since a long gap may indicate a dispreferred response (Kendrick & Torreira, 2015).

In contrast to previous studies showing that eyebrow raises play a role in conveying questions (Bavelas et al., 2014; Borràs-Comes et al., 2014; Borràs-Comes & Prieto, 2011; Chovil, 1991; Coerts, 1992; Cruz et al., 2017; Domaneschi et al., 2017; Ekman, 1979; Nota et al., 2021, 2022; Srinivasan & Massaro, 2003; Torreira & Valtersson, 2015; Zeshan, 2004), there was no facilitation effect for questions with eyebrow raises compared to questions without eyebrow movements. This is in line with our expectation of a stronger facilitation effect for eyebrow frowns compared to eyebrow raises due to the common occurrence of eyebrow raises in other contexts (such as for prosodic emphasis; Swerts & Krahmer, 2008). It could be that this common occurrence made eyebrow raises a weaker signal for any specific social action to such as extent that they were as unhelpful in detecting questions as absent eyebrow movements. Alternatively, it could be that eyebrow frowns are stronger signals for information requests in particular (Nota et al., 2022), which were the questions used in the current study. Eyebrow raises may signal other categories of social actions conveyed through questions. If so, this could have resulted in participants not perceiving eyebrow raises as a reliable signal to identify information requests.

Concerning the control variables question type and utterance duration, our findings are in line with the idea that there are linguistic ways to mark questions in the spoken modality (Levinson, 2013; Slonimska & Roberts, 2017). The fact that interrogative and wh-questions resulted in higher accuracies and faster RTs shows that these typical (front-loaded) lexico-syntactic structures provide strong linguistic cues for rapid question identification. It may be that people benefit less from the visual modality when detecting questions when there are strong linguistic cues, but at least in the present study, we did not find any interaction between eyebrow movement and question type. Future studies may vary linguistic features (including prosody) in conjunction with visual signals more systematically and in more detail to address this issue. Unsurprisingly, a longer question duration made it easier to correctly identify the utterance, and slowed down the RT since the RT was measured from the beginning of the video until the button-press. Since all visual signals (including targets and fillers, with the exception of blinks) occurred before or at the verbal utterance onset (average target onset difference of −536 ms, SD = 601 ms, min = −2307, and max = 0 ms) and the fastest RTs for identifying questions were 910 ms after video onset or 1679 ms after utterance onset, none of the participants reacted before the appearance of visual signals or before the verbal utterance started. However, it may be that participants still used a different strategy to distinguish between questions and statements, which may have led to a speed-accuracy trade-off. For example, one strategy may be to wait until the verbal utterance has ended before pressing a button, in order to increase accuracy.

There are some limitations to the current study. Although the study attempted to be as representative of natural behavior as possible by using multiple avatars and diverse stimuli (with various utterance word lengths, different utterance and facial signal durations, and different question types) while keeping experimental control, the eyebrow movement intensities were fixed. Thus, some signals might have been enhanced when the eyebrow movement intensities were originally occurring with lower intensities in the corpus. Moreover, participants did not interact with the virtual avatars by giving overt responses, since they were instructed to use a button-press. Therefore, the timings of participants' responses may not generalize to real turn-taking timings during actual face-to-face conversation.

The interaction between prosody and facial signals was also not investigated, since we kept the prosody intact, and only manipulated the eyebrow movements between conditions in this study. There may be an interaction between eyebrow movements and prosody. However, we expect that if prosody still had an effect on our results, it would have weakened the facilitative effect of eyebrow frowns, since a strong prosodic cue may potentially lead to a weaker effect of the visual modality. Future studies are needed to evaluate the interaction between prosodic features, visual signals, and intensity variation, in marking questionhood and the effects on early question recognition.

Investigating communication using a natural scenario that is representative of how people communicate in real life always involves a trade-off. The more natural an experiment is, the noisier and less experimental control there usually is, and vice versa. Virtual reality technology allows the creation of environments closer to real-life situations than many other stimuli while maintaining the experimental control required for reliable data collection (Peeters, 2019; Tromp, Peeters, Meyer, & Hagoort, 2018). However, it is a rapidly developing field, and by now more photorealistic virtual avatars are already available. The avatars we used for this study were evaluated to be mediocre in terms of humanness, ease of understanding, and likeability. A more realistic avatar may be more representative of natural human behavior, as long as it does not become too ‘‘uncanny’’ (Mori, 1970; Pan & Hamilton, 2018). So far, research using rudimentary avatars has often found comparable psycholinguistic effects to human–human interaction (Heyselaar et al., 2017a, 2017b), making it likely that the current findings are generalizable (however, see Hayes, Crowell, & Riek, 2013, for a study comparing the processing of cospeech gesture from robots versus humans). Future studies investigating human communication using different approaches and methodologies with different levels of experimental control will help to determine in which circumstances visual behavior actually contributes to communication. It may be that using visual behavior to signal questions is a flexible strategy that depends on the specific virtual communicative setting. Furthermore, investigating whether virtual avatars with different levels of human appearance lead to different results may offer insights on the threshold for perceiving virtual characters as convincing interaction partners.

5. Conclusion

Taken together, the main result of the current study is that eyebrow frowns facilitate the identification of questions, but eyebrow raises do not appear to have the same effect. This suggests that specific facial signals critically contribute to the communication of fundamental social actions, like questions, in face-to-face conversation by signaling social action-specific visual information, and provide visual cues to speakers' intentions. This study demonstrates the important role of visual signaling in human communication, and is especially informative for research that intends to study the cognitive and neural basis of social action recognition in face-to-face human communication. Furthermore, this study may offer important insights for the development of virtual characters, by incorporating early eyebrow movement onsets and long eyebrow frowns to more clearly communicate questionhood.

Open Research Badges

This article has earned Open Data and Preregistered Research Design badges. Data is available at https://osf.io/xjpaq/?view_only=f4d7697e1d32428e99322b52f7bb5d51 and preregistered design is available at https://aspredicted.org/1T9_H8R.

Note

[1] Due to the global pandemic, the Ethics Committee approved non-invasive research to be performed online under conditions that this study met.

References

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. (2020). Online timing accuracy and precision: A comparison of platforms, browsers, and participant's devices. *PsyArXiv*. https://doi.org/10.31234/osf.io/jfeca

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Austin, J. (1962). *How to do things with words*. Oxford: Oxford University Press.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59–66). https://doi.org/10.1109/FG.2018.00019

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. https://doi.org/10.1023/b:jAdd.0000022607.19833.00

Bastioni, M. (2021). *MB-Lab*. https://github.com/animate1978/MB-Lab

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, *19*(2), 163–194. https://doi.org/10.1177/0261927×00019002001

Bavelas, J. B., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture*, *17*(1), 98–127. https://doi.org/10.1075/gest.00012.bav

Bavelas, J. B., Gerwing, J., & Healing, S. (2014). Hand and facial gestures in conversational interaction. In T. Holtgraves (Ed.), *The Oxford handbook of language and social psychology*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199838639.013.008

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no … How the brain interprets the pregnant pause in conversation. *PLoS One*, *10*(12), e0145474. https://doi.org/10.1371/journal.pone.0145474

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2020). Conversational expectations get revised as response latencies unfold. *Language, Cognition and Neuroscience*, *35*(6), 766–779. https://doi.org/10.1080/23273798.2019.1590609

Borràs-Comes, J., Kaland, C., Prieto, P., & Swerts, M. (2014). Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, *38*(1), 53–66. https://doi.org/10.1007/s10919-013-0162-0

Borràs-Comes, J., & Prieto, P. (2011). 'Seeing tunes.' The role of visual gestures in tune interpretation. *Laboratory Phonology*, *2*(2), 355–380. https://doi.org/10.1515/labphon.2011.013

Broersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer*. http://www.praat.org/

Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, *25*(1–4), 163–194. https://doi.org/10.1080/08351819109389361

Coerts, J. (1992). *Nonmanual grammatical markers. An analysis of interrogatives, negations and topicalisations in Sign Language of the Netherlands*. Amsterdam: Universiteit van Amsterdam.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Community, B. O. (2018). *Blender—A 3D modelling and rendering package*. Amsterdam: Stichting Blender Foundation. Retrieved from http://www.blender.org

Crespo Sendra, V., Kaland, C., Swerts, M., & Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, *47*(1), 1–13. https://doi.org/10.1016/j.pragma.2012.08.008

Cruz, M., Swerts, M., & Frota, S. (2017). The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties. *Laboratory Phonology*, *8*(1), 23, 24. https://doi.org/10.5334/labphon.110

Domaneschi, F., Passarelli, M., & Chiorri, C. (2017). Facial expressions and speech acts: Experimental evidences on the role of the upper face as an illocutionary force indicating device in language comprehension. *Cognitive Processing*, *18*(3), 285–306. https://doi.org/10.1007/s10339-017-0809-6

Ekman, P. (1979). *About brows: Emotional and conversational signals. Human ethology* (pp. 163–202). Cambridge: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1978). *Manual of the facial action coding system (FACS)*. Palo Alto: Consulting Psychologists Press.

Englert, C. (2010). Questions and responses in Dutch conversations. *Journal of Pragmatics*, *42*(10), 2666–2684. https://doi.org/10.1016/j.pragma.2010.04.005

Gisladottir, R. S., Bögels, S., & Levinson, S. C. (2018). Oscillatory brain responses reflect anticipation during comprehension of speech acts in spoken dialog. *Frontiers in Human Neuroscience*, *12*, 34. https://doi.org/10.3389/fnhum.2018.00034

Gisladottir, R. S., Chwila, D., Schriefers, H., & Levinson, S. (2012). Speech act recognition in conversation: Experimental evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34). https://escholarship.org/uc/item/94n80472

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS One*, *10*(3), e0120068. https://doi.org/10.1371/journal.pone.0120068

Groen, Y., Fuermaier, A. B., den Heijer, A. E., Tucha, O., & Althaus, M. (2016). De Nederlandse empathie quotiënt (EQ) en systematiseren quotiënt (SQ). *Wetenschappelijk Tijdschrift Autisme*, *15*(2), 73.

Hayes, C. J., Crowell, C. R., & Riek, L. D. (2013). Automatic processing of irrelevant co-speech gestures with human but not robot actors. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 333–340).

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Heyselaar, E., Hagoort, P., & Segaert, K. (2017a). In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior Research Methods*, *49*(1), 46–60. https://doi.org/10.3758/s13428-015-0688-7

Heyselaar, E., Hagoort, P., & Segaert, K. (2017b). How social opinion influences syntactic processing—An investigation using virtual reality. *PLoS One*, *12*(4), e0174405. https://doi.org/10.1371/journal.pone.0174405

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

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

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Hömke, P., Holler, J., & Levinson, S. C. (2019). *The cooperative eyebrow furrow: A facial signal of insufficient understanding in face-to-face interaction*. Doctoral Dissertation, Radboud University, Nijmegen.

Hömke, P., Levinson, S. C., & Holler, J. (2022). Eyebrow movements as signals of communicative problems in human face-to-face interaction. *PsyArXiv*. https://doi.org/10.31234/osf.io/3jnmt

House, D. (2002). Perception of question intonation and facial gestures. *TMH-QPSR Fonetik*, *44*, 41–44.

Kaukomaa, T., Peräkylä, A., & Ruusuvuori, J. (2014). Foreshadowing a problem: Turn-opening frowns in conversation. *Journal of Pragmatics*, *71*, 132–147. https://doi.org/10.1016/j.pragma.2014.08.002

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, *52*, 255–289.

Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/ASR

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. JSTOR. https://doi.org/10.2307/2529310

Lenth, R. (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Levinson, S. C. (2013). Action formation and ascription. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 101–130). Oxford: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118325001.ch6

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130302. https://doi.org/10.1098/rstb.2013.0302

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731. https://doi.org/10.3389/fpsyg.2015.00731

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. (2000). *Language and gesture*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620850

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*, 1551. https://doi.org/10.3758/s13428-020-01514-0

Miranda, L., Swerts, M., Moraes, J., & Rilliard, A. (2021). The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions. *Language and Speech*, *64*(1), 3–23. https://doi.org/10.1177/0023830919898886

Mori, M. (1970). *The uncanny valley: the original essay by Masahiro Mori*. IEEE Spectrum, 6.

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, *11*(8), 1017. https://doi.org/10.3390/brainsci11081017

Nota, N., Trujillo, J. P., & Holler, J. (2022). Specific facial signals associate with categories of social actions through questions. *PsyArXiv*. https://doi.org/10.31234/osf.io/qrhdf

Pan, X., & Hamilton, A. F. de C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395–417. https://doi.org/10.1111/bjop.12290

Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic Bulletin & Review*, *26*(3), 894–900. https://doi.org/10.3758/s13423-019-01571-3

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, *9*, 1109. https://doi.org/10.3389/fpsyg.2018.01109

Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., … del Mar Vanrell, M. (2015). Intonational phonology of Catalan and its dialectal varieties. In S. Frota & P. P. Vives (Eds.), *Intonation in Romance*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oSo/9780199685332.003.0002

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, *6*, 509. https://doi.org/10.3389/978-2-88919-825-2

Rossano, F. (2010). Questioning and responding in Italian. *Journal of Pragmatics*, *42*(10), 2756–2771. https://doi.org/10.1016/j.pragma.2010.04.010

RStudio Team. (2022). *RStudio: Integrated development environment for R*. RStudio, PBC. http://www.rstudio.com/

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 40.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I*. Cambridge: Cambridge University Press.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. London: Cambridge University Press.

Sicoli, M. A., Stivers, T., Enfield, N., & Levinson, S. C. (2015). Marked initial pitch in questions signals marked communicative function. *Language and Speech*, *58*(2), 204–223. https://doi.org/10.1177/0023830914529247

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category—ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Slonimska, A., & Roberts, S. G. (2017). A case for systematic sound symbolism in pragmatics: Universals in wh-words. *Journal of Pragmatics*, *116*, 1–20. https://doi.org/10.1016/j.pragma.2017.04.004

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, *46*(1), 1–22. https://doi.org/10.1177/00238309030460010201

Stivers, T., & Enfield, N. J. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, *42*(10), 2620–2626. https://doi.org/10.1016/j.pragma.2010.04.002

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K. E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*(2), 219–238. https://doi.org/10.1016/j.wocn.2007.05.001

ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. *PsyArXiv*. https://doi.org/10.31234/osf.io/b5zq7

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.

Torreira, F., & Valtersson, E. (2015). Phonetic and visual cues to questionhood in French conversation. *Phonetica*, *72*, 2. https://doi.org/10.1159/000381723

Tromp, J., Peeters, D., Meyer, A. S., & Hagoort, P. (2018). The combined use of virtual reality and EEG to study language processing in naturalistic environments. *Behavior Research Methods*, *50*(2), 862–869. https://doi.org/10.3758/s13428-017-0911-9

Trujillo, J. P., & Holler, J. (2021). The kinematics of social action: Visual signals provide cues for what interlocutors do in conversation. *Brain Sciences*, *11*(8), 996. https://doi.org/10.3390/brainsci11080996

van der Meer, H. A., Sheftel-Simanova, I., Kan, C. C., & Trujillo, J. P. (2021). Translation, cross-cultural adaptation, and validation of a Dutch version of the actions and feelings questionnaire in autistic and neurotypical adults. *Journal of Autism and Developmental Disorders*, *52*, 1771. https://doi.org/10.1007/s10803-021-05082-w

van der Struijk, S., Huang, H.-H., Mirzaei, M. S., & Nishida, T. (2018). FACSvatar: An open source modular framework for real-time FACS based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 159–164). https://doi.org/10.1145/3267851.3267918

Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change*, *26*(3), 387–420. https://doi.org/10.1017/S0954394514000155

Williams, J. H., & Cameron, I. M. (2017). The Actions and feelings questionnaire in autism and typically developed adults. *Journal of Autism and Developmental Disorders*, *47*, 3418–3430. https://doi.org/10.1007/s10803-017-3244-8

Zeshan, U. (2004). Interrogative constructions in signed languages: Crosslinguistic perspectives. *Language*, *80*(1), 7–39. https://doi.org/10.1353/lan.2004.0050

Appendix

A. Participant characteristics

Characteristic	Mean (*SD*)	Minimum and maximum possible score
EQ	45 (11)	0–80
AFQ	34 (6)	0–54
Avatar evaluation	8 (3)	0–15
1) Humanness	2.81 (1.29)	0–5
2) Ease of understanding	2.78 (1.09)	0–5
3) Likeability	2.66 (1.30)	0–5

B. Exploration of the effect of questionnaire scores on accuracy and response time

When assessing whether questionnaire scores affected accuracies and RTs, we found that there was a significant effect of avatar evaluation scores for accuracy ($\chi^2(1) = 6.79, p = .009$). This shows that higher avatar evaluation scores (average per participant for avatar humanness, ease of understanding, and likeability) related to higher accuracies. There was no significant effect of avatar evaluation scores for RT. Furthermore, there was no significant effect of EQ and AFQ scores for accuracy nor for RT.