School of Electronic Engineering and Computer Science
Queen Mary University of London

Head movement in conversation

Tom Gurion

Submitted in partial fulfillment of the requirements
of the Degree of Doctor of Philosophy

August 2021

Statement of originality

I, Tom Gurion, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Abstract

This work explores the function and form of head movement and specifically head nods in free conversation. It opens with a comparison of three theories that are often considered as triggers for head nods: mimicry, backchannel responses, and responses to speakers' trouble. Early in this work it is assumed that head nods are well defined in terms of movement, and that they can be directly attributed, or at least better explained, by one theory compared to the others. To test that, comparisons between the theories are conducted following two different approaches. In one set of experiments a novel virtual reality method enables the analysis of perceived plausibility of head nods generated by models inspired by these theories. The results suggest that participants could not consciously assess differences between the predictions of the different theories. In part, this is due to a mixture of gamification and study design challenges. In addition, these experiments raise the question of whether or not it is reasonable to expect people to consciously process and report issues with the non-verbal behaviour of their conversational partners. In a second set of experiments the predictions of the theories are compared directly to head nods that are automatically detected from motion capture data. Matching the predictions with automatically detected head nods showed that not only are most predictions wrong, but also that most of the detected head nods are not accounted by any of the theories under question. Whereas these experiments do not adequately answer which theory best describe head nods in conversation, they suggest new avenues to explore: are head nods well defined in the sense that multiple people will agree that a specific motion is a head nod? and if so, what are their movement characteristics and what is their reliance on conversational context? Exploring these questions revealed a complex picture of what people consider to be head nods and their reliance on context. First, the agreement on what is a head nod is moderate, even when annotators are presented with video snippets that include only automatically detected nods. Second, head nods share movement characteristics with other behaviours, specifically laughter. Lastly, head nods are more accurately defined by their semantic characteristics than by their movement properties, suggesting that future detectors should incorporate more contextual features than movement alone. Overall, this thesis questions the coherence of our intuitive notion of a head nod and the adequacy of current approaches to describe the movements involved. It shows how some of the common theories that describe head movement and nods fail to explain most head movement in free conversation. In addition, it highlights subtleties in head

movement and nods that are often overlooked. The findings from this work can inform the development of future head nods detection approaches, and provide a better understanding of non-verbal communication in general.

Acknowledgements

First and foremost I would like to thank my partner Mitch, for standing by my side and support me through the ups and downs of the PhD. I could not have done it without you. I would like to thank my supervisors, Pat and Julian for the insightful direction and interesting conversation. Our meetings were always encouraging and uplifting, even when (or perhaps because) I kept doubting my work. I am grateful for the people I worked with in The Cognitive Science Research Group and in The Media and Arts Technology Programme.[1] Many of you became much more than colleagues for me, as we explored what London has to offer together, discovered new and old hobbies and shared interests. Thank you for the hours of playing board games, cycling, music making, bouldering, and more. Last but not least, I would like to thank my family and friends back home for believing in me and being there to hear my endless rants about the PhD.

[1]The EPSRC and AHRC also deserve my gratitude as this work is sponsored by The EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

Contents

Acronyms

ECA embodied conversational agent. 5, 13, 14, 17, 38, 40–42, 53, 54, 66, 79

FCD floor control detection. 6, 22, 24, 26, 28–39, 58, 59, 68, 80

FP faking period. 45, 49–53

LSTM long short-term memory. 27, 28, 30, 31, 33–38

ML machine learning. 6, 13, 14, 38, 39, 43, 67

POI point of interest. 67, 68, 71–75, 78

RMS root mean square. 22, 24, 26, 36, 38

VAD voice activation detection. 22, 23, 26, 39, 45, 46, 58

VR virtual reality. 6–8, 15, 17, 40, 41, 43–49, 53, 54, 56–63, 79, 80

Chapter 1

Introduction

1.1 Motivations and contributions

This thesis explores head movement, and head nods in particular, in free conversation. The literature on head movement and nods is scattered around a few theoretical branches that are rarely compared. For example, evidence suggests that people tend to automatically and unconsciously mimic the head movement of their conversational partners. Another branch of research shows that listeners' responses such as head nods appear when the speaker expects them at the end of their turn, as presentations of acknowledgement or support. These two explanations, among others, can often account for the same observed movement, while also predict behaviours that are fundamentally different from each other. These predictions were rarely compared in the past. One major goal of this thesis is therefore to collect some of the common theories that explain head nods and compare their predictions.

A comparison of the predictions of such theories, in the context of this work, does not attempt to determine if one theory should be considered as a better explanation of head nods than the others. After all, head nods can be attributed to more than one theoretical explanation. Even a single head nod, in some cases, can be attributed to more than one explanation. The comparison here therefore tries to explore how well the theories account for the varied observed head movement in conversation and what are the characteristics of the predictions of each theory when compared to observed behaviour.

Another focus of this work is on the methodology of testing different theories accounting for head movement. Studies of non-verbal communication often rely on confederates or opt to employ embodied conversational agents (ECAs) and restrict the conversation as ways to provide consistent manipulation. These approaches are problematic in regards to the ecological validity of the results.

Furthermore, when the manipulation is assessed with self-reported measures, it is often done with post-experience questionnaires. These are known to be problematic as participants often form a post-hoc rational for their action. To solve these issues, a novel virtual reality (VR) method for studying non-verbal communication is developed. Its goal is to enable the real-time manipulation of non-verbal cues in a shared VR environment. At the same time, it provides means for the participants to rate the plausibility of the non-verbal behaviour of their conversational partners in real time through a gamified method.

Exploring and contrasting theories for head movement highlighted the importance of turn-taking as context for their predictions. When analysing conversations computationally (as oppose to manual annotations), this context is not always easy to extract. Many studies opt to employ sophisticated machine learning (ML) algorithms to detect turn changes. These often lack real-time processing capabilities. For example, they cannot be used to animate the body language of speakers in virtual environments in real time. They are also hard to implement and to integrate into other research software. A solution is suggested in the form of a simple audio based algorithm that can detect turn changes in real time, light on resources, and is simple to implement in almost any programming language or environment.

Last but not least, this work recognises issues with the common definition of head nods as an up and down head movement. These were discovered through a series of studies that assume that head nods can be automatically detected by analysing motion captured data. Their results show discrepancies between the detected nods and the expected behaviour. A closer investigation of the automatically detected nods, in comparison to annotated ones, reveals that what people perceive as a head nod is significantly different from the up and down head movement commonly described in the literature.

1.2 Thesis structure

Chapter 2 reviews the literature and state-of-the-art research that prepare the reader to better understand the context and contributions presented in this thesis.

Chapter 3 presents a simple algorithm for detecting floor control in conversation. State-of-the-art approaches using ML can achieve good accuracy but often lack the incremental, real-time processing needed by various applications. They also require annotated data for training, and due to their black box nature can be tricky to interpret. An alternative floor control detection (FCD) algorithm that uses real-time audio signal pro-

cessing, two main parameters and assumes separate microphones for each participant is presented. Comparison of performance on a corpus of unstructured dialogues shows that the algorithm is more accurate than a random predictor and it performs as well as specially trained state-of-the-art deep learning model. Depending on parameters the algorithm presented here is also more stable and faster than all three comparison algorithms. Additionally, it is language independent, robust to audio bleed between microphones and models backchannel responses appropriately. Potential applications include automatic annotation of dialogue for interaction research, through meeting summaries, to driving animations of speakers and listeners in social VR.

Chapter 4 proposes a VR system for studying non-verbal behaviours, and uses it to compare three common theories that explain head nods in free conversation. The system enables the development of what could be described as partial Turing tests: models of non-verbal behaviour are implanted into the virtual avatars of the participants and override parts of their behaviours on demand. At the same time, the participants are encouraged to detect when the avatar is driven by the algorithm versus driven by their conversational partner's movement. This effectively provides ratings for models of non-verbal cues. The system is used in a study to compare models of head nods in free conversation based on three theories: mimicry, backchannel responses, and listeners' responses to speakers' trouble. The results highlight the advantages and disadvantages of the VR method, and provide insights into gamified study designs in VR. The comparison between the theories, however, is inconclusive. It implies that people usually do not process the non-verbal behaviours of their conversational partner in a conscious way, hence cannot report when these go wrong.

Chapter 5 discusses the comparison of predictions from the three theories mentioned above to movement data in conversation. Two studies are presented, one is conducted with participants conversing face-to-face while the other replicates the same design but in a VR environment. In both studies pairs of participants freely converse for 15 minutes while fitted with a motion capture system. Head nods are automatically extracted from the data, and compared to theoretical predictions. The results highlight the differences between head nod patterns of speakers and listeners. They suggest that the theories produce predictions that overlap above what is expected by chance, suggesting that they are not mutually exclusive. The predictions found to be significantly different from the detected head nods, calling into question the competency of simple methods to de-

tect head nods from movement data. In addition, the vast majority of the observed head nods are not accounted by any of the investigated models, implying that mimicry, backchannel responses, and responses to speech disfluencies cannot explain head nods exhaustively.

Chapter 6 investigates the definition of a head nod. A naive description of a head nod is a downwards movement of the head, followed by an upwards movement. This is inline with the automatic head nods detector employed throughout this thesis. Using a dataset of short video snippets of dialogues two sets of algorithmically detected head movements are extracted; a set of nods and a set of peak head movements. Annotation of these datasets suggests that most algorithmically detected nods are not perceived as nods. Clustering of annotations indicates that speaker and listener head movements are perceived differently and involve a mixture of other behaviours, most notably laughter and posture shifts. It shows that head nods are more accurately defined by their semantic characteristics than by their movement properties, suggesting that future detectors should incorporate more contextual features than movement alone.

Chapter 7 Concludes with a general discussion and suggestions for future work.

1.3 Associated publications

Portions of the work detailed in this thesis have been presented in national and international scholarly publications. Some passages have been quoted verbatim from these sources.

- The VR system discussed in Section 4.2 was published as a short conference paper (Gurion et al., 2018).
- The face-to-face study from Chapter 5 was published as a conference paper (Gurion et al., 2020).

Chapter 2

Literature review of head movement in conversation

This chapter reviews concepts related to head movement in conversation. It touches the subject from a few different angles to explain what behaviours are usually associated with head movement. One prominent example, which is the centre of this thesis, is the idea head nods, often described as the down then up movement of the head. Head nods are discussed in the context of manual annotations and automatic detection. The chapter also highlights some theories related to non-verbal communication in general and head movement and nods in particular that are discussed extensively throughout this work.

2.1 Annotation schemes for head movement

Annotation schemes for head movements can provide insight into the properties that researchers find important. The annotation schemes that are discussed here are those that deal specifically with head movement, and aim to describe the dynamics of the conversation. This is as oppose, for example, to annotation schemes for emotional expression.

The MUMIN annotation scheme (Allwood et al., 2007) is designed to study the function of multimodal communicative expressions, with emphasis on feedback, turn management, and sequencing functions. Table 2.1 presents the attributes and possible values suggested by this annotation scheme. Note in particular the large set of possible head gestures, 12 in total. These include values that can be easily mixed up like `Single Jerk` (which usually means a head nod

Attribute	Possible values
Feedback give	
Basic	Contact & Perception, Contact Perception & Understanding
Acceptance	Accept, Non-accept
Additional Emotion / Attitude	Happy, Sad, Surprised, Disgusted, Angry, Frightened, Certain, Uncertain, Interested, Uninterested, Disappointed, Satisfied, Other
Feedback elicit	
(same attributes and values as Feedback give)	
Turn-management	
Turn-gain	Turn-take, Turn-accept
Turn-end	Turn-yield, Turn-elicit, Turn-complete
Turn-hold	*binary*
Sequencing	
Opening sequence	*binary*
Continue sequence	*binary*
Closing sequence	*binary*
Hand gestures	
Handedness	Both hands, Single hand
Trajectory	Up, Down, Sideways, Complex, Other
Semiotic type	Indexical Deictic, Indexical Non-deictic, Iconic, Symbolic
Communicative function	Feedback give, Feedback elicit, Turn management, Sequencing
Facial display	
General face	Smile, Laughter, Scowl, Other
Eyebrows	Frowning, Raising, Other
Eyes	Exaggerated Opening, Closing-both, Closing-one, Closing-repeated, Other
Gaze	Towards interlocutor, Up, Down, Sideways, Other
Mouth	Open mouth, Close mouth, Corners up, Corners down, Protruded, Retracted
Head	Single Nod, Repeated Nods, Single Jerk, Repeated Jerks, Single Slow Backwards Up, Move Forward, Move Backward, Single Tilt, Repeated Tilts, Side-turn, Shake, Waggle
Semiotic type	(same as Hand gestures)
Communicative function	(same as Hand gestures)
Multimodal relations	
Cross-modal function	Non-dependent, Dependent-compatible, Dependent-incompatible

Table 2.1: Attributes and possible values of the MUMIN annotation scheme (Allwood et al., 2007).

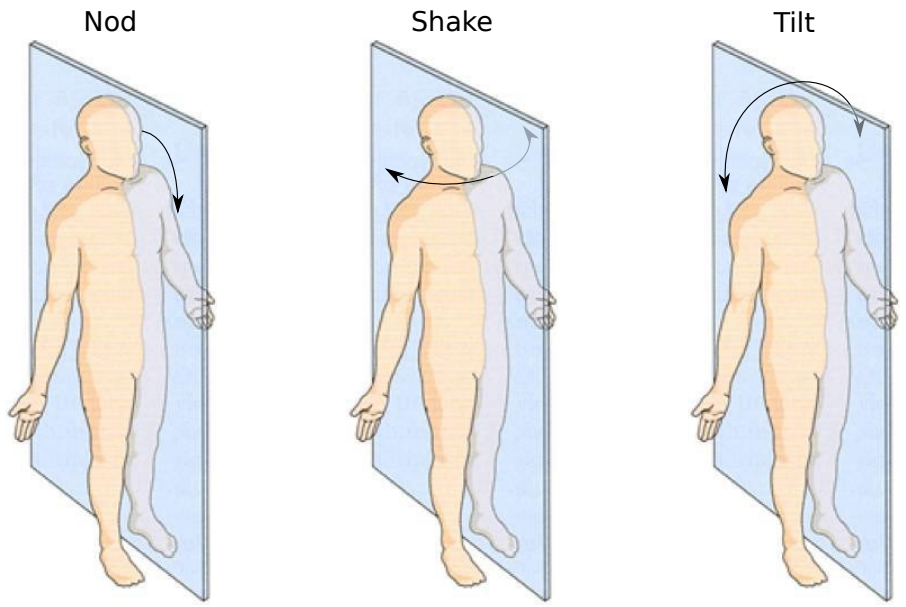


Figure 2.1: Canonic examples of movement associated with head nods, shakes, and tilts.

that starts with a significant upwards stroke), and `Single Slow Backwards Up`.

Three case studies demonstrate the application of the MUMIN annotation scheme to video captured conversations (Allwood et al., 2007). In the first case study two annotators annotated one minute of an interview of an actress for Danish television. The annotators mostly agreed on the selection of "communicative gestures": the first annotator selected 37 gestures, while the second 33 gestures, of these 29 overlapped. Inter-annotator agreement was calculated by Cohen's kappa. The annotators agreed (kappa >0.6) on the feedback (give and elicit), turn-management, and on all of the facial display attributes but `Head` and `Gaze`. They disagreed (kappa <0.6) on sequencing attributes, particularly on the `Continue sequence` binary attribute. Inter-annotator agreement for hand gestures annotations are not reported, although it is implied that the data contains only facial displays and no hand gestures. Similarly, agreement on multimodal relations is not reported. In the second case study one annotator annotated a one minute video of a dialogue from a Swedish film. In the last case study a one minute video of a television interview is annotated by two pairs of naive coders (who worked together to produce one annotations file), and one expert coder. This is equivalent to three annotators. `Handedness`, `Trajectory`, and `Semiotic type` are reported for the hand gestures, from which only `Handedness` showed agreement (kappa >0.6). Note that none of these case studies represent un-

structured conversation well. This is because they are at least partially scripted and with interactants taking asymmetric roles.

Various studies rely or extend the MUMIN annotation scheme (Navarretta, 2011; Paggio and Navarretta, 2012; Boholm and Allwood, 2010). Common to all of these are the recognition of head nods, shakes, and tilts. Canonical examples for these are illustrated in Figure 2.1.

Having a large set of descriptors for head movement is not unique to the MUMIN annotation scheme. `Nod`, `Backnod`, `Double nod`, `Shake`, `Upstroke`, `Downstroke`, `Tilt`, `Turn`, `Waggle`, `Sidenod`, `Backswipe`, and `Sideswipe` is the list of possible head movement annotations suggested by de Kok (2013). Similarly large number of values is available for other annotation schemes (e.g. Kousidis et al., 2013; Blache et al., 2009; Karpiński et al., 2015). Common to all these are nods, jerks (upwards nods), tilts, shakes, and waggles. With the exception of waggles (defined as irregular connected movement by Kousidis et al., 2013), these can be described as either vertical or horizontal movement. Some schemes also record single versus multiple repetitions (Boholm and Allwood, 2010; Blache et al., 2009), and the initial direction of the movement (Boholm and Allwood, 2010; de Kok, 2013).

A common trend in the literature is to first ask annotators to mark regions with movements that present a communicative function or response (Allwood et al., 2007; Kousidis et al., 2013; de Kok, 2013). These are not defined by the annotation scheme, leaving room for annotators to choose which movements require annotation and which do not. de Kok (2013) describes this as follows: "First the interesting regions with listener responses were identified. This was done by looking at the video of the listener with sound of the speaker and marking moments in which a response of the listener to the speaker was noticed. In the second step these regions were annotated ... In the third and final step the onset of the response was determined."

It is important to note that most of the annotation schemes discussed above are evaluated with participants that are unfamiliar to each other (Boholm and Allwood, 2010; Paggio and Navarretta, 2012). Whereas this is not a problem by itself, it might explain differences in approach to most of the work presented here, as the participants in most of the studies in this thesis are familiar with each other.

Lastly, evaluations of the annotation scheme often rely on map tasks (Anderson et al., 1991), or otherwise structured or asymmetric conversations (Paggio and Navarretta, 2010; de Kok, 2013). This poses a question for how well these annotations schemes capture movement and semantics in free conversation.

2.2 Automatic head nod detection

Automatic head nod detection techniques often appear in studies exploring the development of embodied conversational agents (ECAs), as these should be able to process verbal and non-verbal communication. Alternatively, these techniques enable the automatic annotations of datasets, making research of large corpora possible.

Head nod detection is often based on specialised motion capture equipment (Cerrato and Svanfeldt, 2006; Healey et al., 2013; Chen et al., 2015), or on processing of video data with computer vision (Kawato and Ohya, 2000; Kapoor and Picard, 2001; Kang et al., 2006; Dong et al., 2007; Nguyen et al., 2012). One advantage of motion capture approaches is that both 3 dimensional position and rotation information are available. Computer vision approaches, on the other hand, are easier to set up and are less obtrusive. They usually detect and track pupils or face area for horizontal and vertical movement. The data from either system is then processed by either a set of rules or machine learning (ML) algorithms to detect head nods and other head gestures.

An early example of a computer vision rule-based detector is suggested by Kawato and Ohya (2000). It tracks the point between the eyes at 13 frames per second, while the resolution sets the distance between the eyes at approximately 14 pixels. Each frame is defined as `stable`, `extreme`, or `transient`, based on the vertical position of the tracked point. If the 5 frames around the frame in question have the point within 2 pixels from each other, the frame is `stable`. If the point is the highest or lowest among the 5 frames around it the frame is an `extreme`. Otherwise the frame is `transient`. If there are at least three `extreme` frames with no stable frames between them, and adjacent `extreme` values differ by at least 3 pixels, the system reports a head nod. This system is evaluated against videos of participants instructed to move their heads up and down or left and right, raising the question of the ecological validity of the results. The results include average accuracy of 92% with recall of 94%.

Healey et al. (2013) suggest a similar approach to process the head height information produced by a motion capture system. Their method is employed extensively throughout this work, and is described in detailed in Section 4.1.1.

ML approaches employ a variety of models like Hidden Markov Models and Support Vector Machines to predict head nods from features like the head position, velocity, and frequency (Kapoor and Picard, 2001; Morency et al., 2005; Kang et al., 2006; Nguyen et al., 2012; Chen et al., 2015). An early ML detector tracks the pupils with an infrared camera and process these with a Hidden Markov Model to detect head nods and head shakes in real time (Kapoor and Picard, 2001). Similarly to Kawato and Ohya (2000), the dataset for training

their model, as well as for evaluating it, is collected by asking participants to answer yes or no questions only by head nods or head shakes. Although this dataset contains head nods, it is expected to be significantly different from head movement found in free conversation. For example, head nods as backchannel responses (discussed next) are not necessarily an answer to a question equivalent to 'yes'. The evaluation suggests that the recall of the system for head nods is 92%, and for head shakes is 95%. The accuracy in both cases is perfect (no false detections).

Many of these studies, however, do not suffer from this ecological validity issue. More recent works often evaluate their detectors on free conversations with annotated head nods (Nguyen et al., 2012; Chen et al., 2015).

Note that some of these detectors are able to detect more than just head nods. In that case, they usually detect nods from vertical movements and shakes from horizontal movements (Kawato and Ohya, 2000; Morency et al., 2005), suggesting that these are the most basic division of head movement into two distinct classes. This is inline with the common attributes suggested by annotation schemes for head movements as discussed earlier in Section 2.1.

Whereas most studies process only head movement to determine head nods, evidence suggests that incorporating conversational context improves detection. This has been demonstrated by adding the speaker's speaking state (Nguyen et al., 2012), or information from the internal dialogue manager of an ECA (Morency et al., 2005) as features to ML based head nods detectors.

Note also that some of the studies above analyse features that are based on head height (or pupils height as a proxy; Kawato and Ohya 2000; Healey et al. 2013). Other studies consider the pitch — the rotation of the head around an axis parallel to an imaginary line between the ears — as the basic feature for detection (Chen and Harper, 2009).

2.3 Head movement mimicry

Conversational partners often mimic each other' postures, gestures, use of language, facial expression, and more. Evidence suggests that mimicry in conversation has a persuasive power and can increase likeability, empathy, and feeling of closeness (for a review see Chartrand and Lakin, 2013).

Chartrand and Bargh (1999) found an automatic tendency to imitate others in social interactions. They call this mimicry behaviour the "chameleon effect". In three experiments participants interacted with confederates. In the first experiment the confederates either rubbed their face or shook their foot, and either smiled or not. Naive judges coded the video recorded interactions. Their coding indicates significant mimicry of both facial expression and behaviour,

and that behaviour matching (face rubbing) was independent of facial expression (smiling). In the second experiment the confederate either mimicked the participant or not. After the experiment, the participants self-reported higher levels of rapport for the mimicking confederates. In both studies, when asked, the participants were unaware of the mimicry. The third experiment shows correlation between perspective-taking — the ability to imagine yourself in the shoes of someone else — and mimicry. The authors explain these results by the perception-behaviour link: a cognitive mechanism that links together perception and action and suggests that we have to act in order to perceive. This theoretical reasoning suggests that mimicry is an unconscious automatic behaviour that is a by product of our ability to process social encounters.

Bailenson and Yee (2005) extend these ideas to virtual reality (VR). Using a virtual agent they proposed a more concrete claim. They conducted an experiment with participants telling a story to an agent, who either mimics their head movement in a 4 seconds delay, or plays back a recorded movement from an interaction with previous participants. Participants in the mimicry condition reported higher levels of rapport. Using similar methods, later studies show that there is a high correlation between social anxiety and lack of mimicry (Vrijsen et al., 2010a,b); We are more open to outgroup members if they mimic (Hasler et al., 2014); And that mimicry increases trust (Verberne et al., 2013).

The amount of mimicry we apply depends on context. When affiliation is an explicit goal people tend to mimic each other more, and formal interactions involve less mimicry than friendly encounters (Lakin and Chartrand, 2003). In one study an experimenter was either friendly and informal or polite and highly professional, and either mimicked the participant gesture, posture, and mannerisms or not. Participants reported that they felt physically colder when the experimenter was friendly and non-mimicking or polite and mimicking (Leander et al., 2012). Mimicry also depends on the type of the interaction. In a competitive investment game, mimicking virtual agents did not increase neither liking nor trust, whereas in a cooperative route planing game, the mimicking agent was rated as more likeable and trustworthy (Verberne et al., 2013).

Some studies, however, provide evidence in the opposite direction to the chameleon effect. Tiedens and Fragale (2003) compare the claims about mimicry with a hypothesis that people compliment each others posture during interaction: when one person pose is expanded, the other will opt for a submissive, constricted stance, and vice versa. They conducted studies using confederates that presented expanded, natural, or constricted pose. Participants tended to present the opposite posture to the confederate, supporting the researchers hypothesis against the common mimicry claim. Furthermore, when the participants were asked whether they liked the confederate and felt comfortable with

them, their answers were more positive if they presented an opposite posture to the confederate.

Recently, a study that explored the effect of mimicry on rapport found no conclusive evidence (Hale and Hamilton, 2016). In their first experiment, participants interact with a virtual agent that either mimics their head and torso movement or not, in a 1 or 3 seconds delay. Participants reported higher level of rapport to the mimicking agents in the 3 seconds condition. Other self-reported measures, like trust and similarity to the agent, and the general smoothness of the interaction, were not found to be affected by mimicry. In their second experiment East Asian and European participants interacted with East Asian or European agents, that either mimicked or not in 3 seconds delay. The researchers hypothesised that in addition to an increased rapport for the mimicry condition, participants rapport will also be higher for agents with matching ethnicity. Surprisingly, neither effect was found, not on rapport, nor on any of the other self-reported measures.

Hale et al. (2019) recently challenged the mimicry time lag that many of the studies above report. The common 4 seconds delay claim from Bailenson's work (Bailenson et al., 2004; Bailenson and Yee, 2005) suggests that mimicry is not a reactive process and it relies on memory to operate. This seemed rather unlikely to the authors, so they tested two alternative hypotheses. One alternative model describes mimicry as a predictive process, similar the mechanism that allows musicians to coordinate their playing, and implies no time lag. The other model describes mimicry as an immediate reactive response to other people, thus implies a time lag on the order of 300-1000 milliseconds. They used motion capture technologies to measure the head pitch in dyadic conversations and found that listeners low-frequency head movement follows the speaker's head movement in a 600ms delay. Another finding from the same study is that there is a negative correlation between speakers and listeners head movements in higher frequencies. More specifically, listeners often move their head in 2.6–6.5Hz while speakers rarely do so.

Most of the literature about head movement and mimicry does not specifically deal with head nods. When tested, researchers found that head nodding mimicry correlates to ones overall communication skill (Wu et al., 2020). A study with medical students and simulated patients shows that when the medical student mimics the head nods of the patient they receive a higher score on their post consultation survey, as reported by the patient.

Provide a backchannel feedback upon detection of:
P1 a region of pitch less than the 26th-percentile pitch level and
P2 continuing for at least 110ms,
P3 coming after at least 700ms of speech,
P4 provided that no backchannel has been output within the preceding 800ms,
P5 after 700ms wait.

Table 2.2: Prosody based algorithm to predict backchannel responses (Ward and Tsukahara 2000; copied from Poppe et al. 2010).

2.4 Head nods as backchannel responses

Natural and engaging conversations rely on the ability of the interlocutors to achieve common ground. This process often involve feedback from listeners for their understanding, or misunderstanding. Timed listeners responses, known as backchannel responses (Yngve, 1970), can be non-verbal (e.g. head nods), or paralingual (e.g. utterances like "uh-huh"). They are crucial for the speaker to assess the listener engagement and adapt to it (Bavelas et al., 2000), and can mediate turn-taking in conversation (Duncan et al., 1979).

Modelling backchannel responses is an increasingly relevant topic for research. In addition to their contribution to our understanding of the cognitive processes behind social interaction, they see applications in a wide array of fields. Non-verbal and paralingual models can be implemented to make ECAs more realistic, and may support new telecommunication technologies like social VR.

Surface features like speaker-listener eye contact, speaker voice level, and prosody are often enough to model backchannel responses. Ward and Tsukahara (2000) suggest a prediction model that is based on the speaker prosody. Their model, summarized in Table 2.2, provides a set of simple hand-crafted rules that use the speaker "vocalisation state" (speaking or silent) and pitch, to predict listeners' backchannel responses. Many studies relay on this model, either as a component for conversational systems (Maatman et al., 2005), or as a reference point for an alternative, sometimes more complex, predictor (Morency et al., 2008; Poppe et al., 2010).

Other models incorporate the speaker-listener eye-contact (Poppe et al., 2010). These models are influenced by research of gaze in conversation (Goodwin, 1979) and experimental studies that found relationships between backchannel responses and mutual gaze (Bavelas et al., 2002). Other approaches suggest using the speaker head movements (Maatman et al., 2005; Gratch et al., 2006), or surface text features (Lee and Marsella, 2006).

Whereas all of these studies use a set of carefully chosen rules to predict backchannel responses, recent studies suggest data-driven approaches. Nishimura

et al. (2007) predict listener responses using the pitch and power of the speaker voice. Their model, however, isn't constrained to predict only backchannel responses. A response generator uses audio features to choose between a backchannel response, collaborative completions (e.g. suggesting a keyword to the speaker), and other types of response. Their decision-tree model was trained using audio-recorded and annotated conversations of the RWC multimodal dataset (Hayamizu et al., 1996). Morency et al. (2008) followed a similar path but with a more exhaustive set of audio-based features, and speaker-listener gaze annotations. In addition, they encoded each feature with a set of encoding templates that modified the appearance of the feature over time. Then, they applied probabilistic methods to choose the feature with the most predictive power to incorporate into the model. Huang et al. (2011) used a simplified model that is based only on non-verbal information. Their model uses the speaker vocalisation state, speaker-listener eye contact, and, interestingly, the speaker smile, to predict listeners' backchannel responses. A year later the model was improved to make it more robust to different speaker behaviours (de Kok and Heylen, 2012). Note that there is almost no direct comparison between the different models, mainly due to the use of different datasets and different evaluation methods (Morency et al., 2008).

2.5 Listeners' responses to speakers' troubles

The interactive nature of face-to-face conversations suggests that all of the participants in a conversation involve in the effort to keep it going. Therefore, listeners responses are especially important when the speaker encounters problems in producing a turn.

Disfluencies are often described as a 3-part structure of reparandom, interregnum, and repair (the terms were proposed by Shriberg, 1994). This structure is presented in the example bellow.

Healey et al. (2013) searched for listeners' backchannel responses after disfluent utterances. They invited groups of 3 participants to discuss an ethical dilemma in a motion captured space. The interaction was manually transcribed to extract disfluencies and head nods were detected automatically from the motion capture data. Their results show that addressees tend to nod in the 1 to 3 seconds window after the disfluency. Similar results were found for side-participants, although with less prominent effect. Moreover, unlike addressees,

side-participants usually suppress backchannel responses as the speaking rate increases. A later study added that upon speech disfluencies listeners gesture more, and with gestures that are similar to the speaker gestures (Healey et al., 2015).

Backchannel responses, however, are not the only possible listener response to speech disfluencies. Listeners often shift their posture when the speaker encounters problems in speech. This idea was proposed as one of many mechanisms behind a multi-modal conversational systems (Maatman et al., 2005). To my knowledge, no systematic evaluation of this claim was done to date.

More generally, it is important to note that disfluencies and repair are an integral part of face-to-face conversations (Schegloff, 1992). Surprisingly, increase in disfluencies and repairs in conversations often predicts less misunderstanding, and not more (McCabe et al., 2013). These properties alone were enough to encourage researchers to develop disfluency detection algorithms, mainly to enable conversational systems to process disfluencies properly.

Various techniques for disfluency detection were suggested throughout the years. They can be generally classified by considering these 3 characteristics: (a) using prosody or other low-level features versus text transcriptions; (b) relying on simple set of rules or machine-learning algorithms; (c) operating in real time (sometime referred to as incremental) or not. Early real-time disfluency detectors relied mainly on low level features, especially prosody (Shriberg et al., 1997; Maatman et al., 2005). Accurate real-time speech-to-text solutions became available in recent years (Hough and Schlangen, 2017). They open the possibility to create transcript based disfluency detectors for real-time operation. One such detector is the "deep disfluency tagger" by Hough and Schlangen (2017). They trained a deep-learning algorithm to tag disfluencies in natural speech. Edit terms like "uh" and "I mean" were tagged with `<e>`, the start of the repair was marked with `<rpS>` (stands for 'repair start'), and the end of the repair with `<rpE>`. The tagger's performance on the disfluency-annotated Switchboard dataset (Meteer, 1995) of phone conversations, is competitive with state-of-the-art models on like-for-like disfluency detection comparison and is the only known system to be usable on and evaluated on automatic speech recognition results. In evaluation on transcripts, the F1 scores reached 0.92 for `<e>` tags and 0.72 for `<rpS>` tags on a per-word basis. On automatic speech recognition results these two F1 scores reached 0.73 and 0.56 using 10-second time windows.

2.6 Directions for research based on the literature

From the literature reviewed above, what is missing in terms of comprehensive real-time testing of models of head nods in dialogue are the following:

- A simulation environment to allow for truly interactive testing of models of nods in free conversation (Chapter 4).
- A turn-taking model for dialogue usable for real-time testing of models of head movement that interact with turn taking structure (Chapter 3).
- A comparison of real-time predictive models of head nods in conversation (Chapter 5).
- A critical evaluation of what underlies theories of head nods and automatic methods of detection in relation to other communicative behaviours (Chapters 5 and 6).

Chapter 3

A simple algorithm for floor control detection in dialogue

This chapter suggests a simple algorithm for detecting floor control in conversation. The performance of the algorithm on a corpus of unstructured dialogue is evaluated and compared to random baseline and a specially trained state-of-the-art deep learning model. This algorithm is used extensively in the following chapters of this thesis to derive, in the abscence of manual annotations, what would be considered the ground-truth for floor control.

The first systematic attempt to describe the organisation of turn-taking in conversation was provided by Sacks, Schegloff and Jefferson (1974). Starting from the observation that people rarely talk over each other they proposed a set of rules for turn-taking that apply independently of the number of people talking, the topic of the conversation, and the length of any given utterance. These rules appear to be universal across languages and provide a critical part of the basic infrastructure of adult conversation and of language acquisition (Stivers et al., 2009; Holler et al., 2016).

The turn-taking rules proposed by Sacks et al. are normative not syntactic; they provide a framework in which the significance of different kinds of silence or overlap between turns can be interpreted. Nonetheless, there are many potential applications for computational systems that are able to track turn-taking in conversation. One simple practical application is to automatically annotate turns in a dataset, a task that is often done manually. These annotations are

essential for calculating turn lengths, relative number of contributions by different participants, patterns of feedback and interruption, and many other basic features of interaction. Automatic turn-detection can reduce costs and effort, and make the annotation of large interaction datasets a practical possibility. Automatic detection of turn changes is also a pre-requisite for the development of more complex applications. For example, modelling speakers' and listeners' non-verbal behaviour (head nods, gestures, etc.) depends on the ability to track who is speaking to whom. Telecommunication technologies can also benefit from such algorithms if they can process data in real time. Social virtual reality, for example, offers the opportunity to simulate and augment aspects of non-verbal communication, including eye gaze, facial expressions, and finger movements. Doing this effectively requires real-time monitoring of turn-taking to allow generation of appropriate non-verbal cues.

The current state-of-the art for modelling turn-taking often employs relatively complex models that require extensive training, particularly deep learning approaches (e.g. Skantze, 2017), and are not always compatible with incremental and real-time processing of dialogue (Anguera et al., 2012). This limits their range of potential applications.

In this chapter, a simple algorithm for floor control detection (FCD) is presented and compared to a state-of-the-art deep learning approach. The algorithm is designed to detect the floor-holder[1] from audio data in natural conversations, in real time. It assumes audio coming from a separate microphone for each interactant and utilises simple signal processing building blocks: root mean square (RMS) calculation, filters, and a threshold. As a result it is relatively simple to implement and adapt to many programming languages and environments. The algorithm is incremental in that it uses only past data for decision making. It is a real-time algorithm in that it is computationally lightweight, hence does not accumulate lag upon operation, and does not require additional resources such as transcription or part-of-speech tagging to operate (like in Skantze, 2017).

3.1 Background

One of the simplest approaches to detect floor control is voice activation detection (VAD) algorithms which aim to detect when someone is speaking. They are used extensively in telecommunication, especially to minimise networking by releasing communication bandwidth for non-speech audio (Benyassine et al., 1997). A limitation of VAD classification is that it is usually too granular,

[1]Some authors distinguish between holding a turn and holding the floor (e.g. Edelsky, 1981). This distinction is suppressed for the purposes of this work.

reporting alternating speech and non-speech segments at a much higher rate than turns in conversation (Ivanov and Riccardi, 2010). Designed primarily for telecommunication, VAD solutions are also prone to audio bleed between microphones in studies that are conducted with all participants in the same room (Bertrand and Moonen, 2010).

Spoken dialogue systems aim to avoid overlap with the user and eliminate unintended long pauses. In general, they try to follow turn-taking behaviour that is similar to human-human interaction, who aim to minimize both overlaps and gaps (Sacks et al., 1974). Early research in turn-taking for spoken dialogue systems attempted to predict turn changes at lexical or time threshold boundaries, by deciding if the dialogue system should take the turn (*shift*) or wait for the user to continue speaking (*hold*) (Maier et al., 2017; Lala et al., 2018). This decision is often based on speech prosody before the boundary, gap duration, and lexical information (Ferrer et al., 2002; Schlangen, 2006).

An alternative approach is to continuously predict the possibility of a turn change (Skantze, 2017; Lala et al., 2019). Skantze, for example, proposes a turn-taking model that continuously predicts the probability that an interactant will speak (Skantze, 2017). The probabilities generated by the models are then used to predict the floor holder, by picking the interactant with the higher probability to speak. This is tested as follows: segments ending with a pause, followed by one and only one speaker in the following one second, are extracted from a dataset. The task for the model is to predict who will be the next speaker after the pause, equivalent to classifying each end of segment as a turn shift or turn hold. The model generates a probability for each interactant separately for 1 second following the pause. The interactant with the higher probability is predicted to be the next floor holder. The performance of the model on this task is significantly better than both a random baseline and human judges facing the same task. Skantze's model is replicated for this chapter as a principal point of comparison.

Speaker diarization aims to answer the question "who spoke when?" (Anguera et al., 2012). It usually deals with unknown number of speakers and one audio source, and attempts to cluster the audio into segments and label the segments by speakers. A few studies in the field suggests that they can achieve 'real-time' performance (e.g. Huang et al., 2007), in the sense that the time it takes to analyse an audio source is shorter than its duration. This is different from the notion of real-time processing discussed in this chapter. From this perspective existing speaker diarization methods are not incremental and do not process input data in real time.

3.2 Floor control definition

Turn-taking is considered in terms of the transitions of turn ownership or *floor control*. The task is posed as identifying, in real time, which conversation partner holds the floor at any given moment, even if there are some concurrent utterances from other interactants, or silence. The terms "turns" and "floor control" are used interchangeably throughout this work. An interactant that has the turn, or controls or holds the floor, is granted the permission to speak by their conversational partners.

There is no one common way to define or annotate turns or floor control. Studies in the field of speaker diarization often follow a similar approach as the one described here: they are interested in segmenting time so that one interactant is considered the speaker at any given moment. Another approach is to define turns as continuous segments of speech. This way, multiple interactants can have overlapping turns, or in the case of long pauses, no-one has the turn (e.g. Ivanov and Riccardi, 2010).

A floor control definition poses some interesting implications for the categorisation of verbal listener responses like "uh-uh" and "yeah", also known as backchannel responses (Yngve, 1970). These are expected by speakers and are a crucial component in conversation (Bavelas et al., 2000). In some cases they are categorised as turns (e.g. Roddy et al., 2018) or changes of floor. This work takes the opposite view: these listener specific behaviours do not indicate a turn, or claim for the the floor. If floor control grants an interactant the permission to speak, a listener response demonstrates that the interactant does not currently hold the floor. This is especially important when the results of a FCD algorithm are used to understand or generate non-verbal behaviour.

3.3 A simple floor control detection algorithm

The proposed FCD algorithm is illustrated in Figure 3.1. It processes buffers of 20ms of audio at a time, arriving from each interactant's microphone. The power of the audio buffers is measured by RMS. The RMS values are passed through a second order low-pass Butterworth filter with a cut-off frequency f. Lastly, the filtered RMS values are compared as follows: If the ratio between the higher and the lower filtered RMS values is greater than a comparison threshold $1 + k$, then the interactant with the higher value is reported as the floor-holder. Otherwise, the floor-holder detected in the previous buffer is reported again, indicating they continue to hold the floor.

Default values for the parameters of this algorithm are manually picked by testing the algorithm performance on one dialogue from the dataset used for

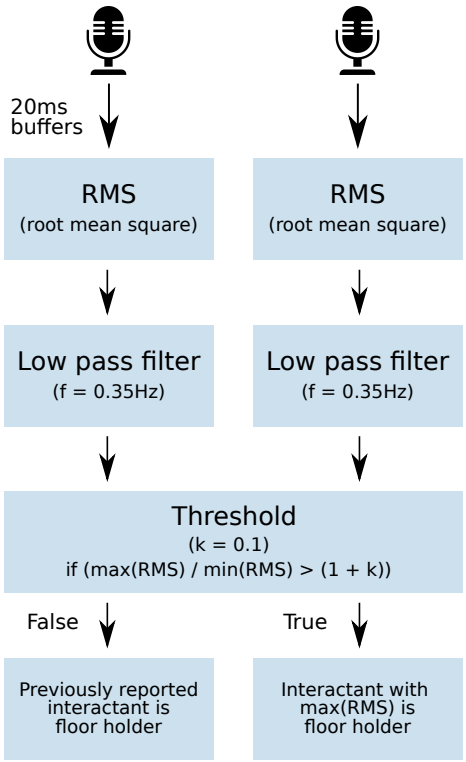


Figure 3.1: Schematic diagram of the floor control detection algorithm.

the evaluation (detailed below), along with data from one session from the face-to-face study described in Section 5.1. The filters' cut-off frequency f is set to 0.35Hz and the comparison threshold k to 0.1. The filters with their low cut-off frequency are used to slow down the changes in the RMS values and were introduced to prevent rapid changes in the resulted floor. The threshold k is chosen to prevent repeated floor changes when the interactants have a similar speech level. This is especially noticeable when there is silence.

During initial testing a variation with VAD instead of the RMS calculation was tried. In dialogues with no audio bleed between the two interactants' microphones the VAD based algorithm performed well. However, cases of audio bleed with speech data in the audio of one interactant, even when only their partner was speaking affected its accuracy. According to this initial investigation using RMS is more robust when audio bleed is present.

A reference implementation for the FCD algorithm, written in Python, and the code for reproducing the evaluation below, are publicly available on GitHub.[2]

3.4 Evaluation

3.4.1 Data

The German subset of the DUEL dataset (Hough et al., 2016) is used for this evaluation. It consists of 30 conversations of 5-20 minutes in length where there are 10 conversations of each different task type, where a task type consists of the following different prompts before the dialogue: (i) discuss a dream apartment that the participants could share (ii) collaboratively design a funny television sketch and (iii) do a role play of a border immigration interview where one participant plays the role of a hostile border force agent and one plays the part of an immigrant seeking entry to a fictional country. In the German subset there are 19 different dialogue pairs and 38 different participants over these 30 conversations. The dataset provides annotations of utterances in Praat TextGrids (Boersma et al., 2002). Each utterance is segmented according to the DUEL project manual.[3] An utterance is segmented into what Meteer (1995) called a 'slash-unit', which corresponds roughly to the grammatical unit of a sentence, but can be below this if the contribution is incomplete.

The dataset also includes audio recordings with one participant on each channel of a stereo audio file. The participants are co-located in the same room during the dialogues, so there is some inevitable audio bleed between their

[2] `https://github.com/Nagasaki45/floor-control/tree/phd-submission`.
[3] `http://www.dsg-bielefeld.de/DUEL/resources/DUEL_manual.pdf`.

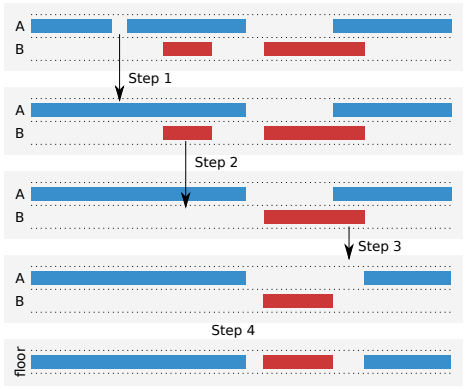


Figure 3.2: Demonstration of calculating the annotated floor from the utterances that are annotated in the dataset, in 4 steps. The horizontal axis represent time. Blue and red rectangles represent intervals of annotated utterances by interactants A and B. The five blocks, one below each other, show the process and intermediate states.

microphones. Manual investigation of one of the dialogues suggests that the bleed is about 20–25dB from one channel to the other. In others words, the voice of one participant can be heard on the opposite channel at about 20–25dB lower level.

The 30 interactions are divided into a training and test sets. Twenty two interactions are in the training set and eight are in the test set (approximately 75% and 25%) where the performance of the competing algorithms is compared. This division is largely for the purposes of training the long short-term memory (LSTM) model (see below), which requires substantial amounts of training data to converge. The results reported below are computed on the test set.

3.4.2 Ground-truth

The DUEL dataset does not include annotations for floor control matching the definition described in Section 3.2. The ground-truth floor is therefore derived automatically from the utterance annotations in the dataset. Figure 3.2 demonstrates how the ground-truth floor is calculated in four steps of transforming the original utterance annotations:

1. Consecutive utterances by the same interactant without interference from their partner are merged together.

2. When interactant A has an utterance that is completely within interactant B's utterance it is deleted, and vice-versa.

3. Remaining overlap annotations are removed.

4. The remaining utterances, free from overlaps, are used as the annotations of who holds the floor.

As implied by Figure 3.2, short segments in the annotated floor remain ambiguous. These segments, involving competition for the floor or silence, cannot be automatically classified based solely on the annotated utterances without introducing more assumptions. Also note that on average, for the entire dataset, competition for floor or silence happen only 16.2% of the time.

Backchannel responses are timed listeners responses like "uh-huh" and head nods (Yngve, 1970). As indicated by their name, verbal backchannel responses accompany the speaker's turn. As such, these listener responses should not occur when holding the floor. In order to analyse backchannel responses correctly the ground-truth for backchannels is extracted automatically from the dataset as follows: an utterance is a backchannel if it is shorter than 500ms and consists only of 'ja', 'okay', 'ohm', 'mhm', and 'genau'. This list of words was suggested by a native German speaker after investigating the audio and the annotations of one of the dialogues in the dataset.

3.4.3 Competitor algorithms

Three competitor algorithms demonstrate how the FCD algorithm compares to a state-of-the-art deep learning turn-taking model and to chance performance estimated by a pseudo random algorithm.

LSTM

This is a replication of the state-of-the-art, LSTM deep learning model of turn-taking suggested by Skantze (2017). The model aims to continuously predict the probability that an interactant will vocalise in the current frame and determine the floor-holder as the interactant with the maximal value.

Skantze discusses two models in their paper. One that is based only on prosodic features, and a full model that takes lexical features (part-of-speech) into account. The replication here is of the prosody model. The prosody model extracts local features from 50ms buffers of audio from the two interactants, and feed these into an LSTM to predict the probability that an interactant will vocalise at the following 60 frames, equivalent to the upcoming 3 seconds.

Below are the implementation details of this replication, including some specific points of difference with Skantze's original model. At the front end, buffers of 50ms of audio are analysed for the following features:

Voice activity: This binary feature does not use audio as an input. It is based on the ground-truth annotations instead. If the interactant is in the middle

of an utterance the voice activity is `True`, `False` otherwise. As discussed below, the use of ground-truth values puts it at an advantage over the FCD model.

Absolute pitch: The fundamental frequency f of the sound in the buffer is calculated using the YIN algorithm (De Cheveigné and Kawahara, 2002).[4] Absolute pitch, in semitones[5], is equal to

$$69 + log_2(f/440)$$

Relative pitch: The absolute pitch is z-normalised per individual.

Voiced: This binary feature is under-specified in the original paper and its description is included with the description of the pitch features. Aiming to keep the features similar to the original paper, the voiced feature is set to `True` when the YIN algorithm reports a valid frequency value and `False` otherwise.

Power: This is the power (intensity) of the signal in decibels, calculated as

$$10 * \log_{10}(\sqrt{\frac{x_0^2 + ... + x_N^2}{N}})$$

where $x_0 ... x_N$ are the buffers' samples. The result is z-normalized per individual.

Spectral flux: In the original paper spectral stability is used, but it is unclear how the fast Fourier transform analysis is divided into bins. To overcome this spectral flux is used instead (Giannoulis et al., 2013), with window size and hop size equal to the entire buffer length (i.e. no overlap between windows). The results are z-normalised per individual.

Two models are trained separately for each speaker. Each one predicts the probability of one interactant's vocalisation independently of the other. The inputs for the models are sequences of 10 seconds of the 12 features listed above (6 for each interactant). In the original paper the author used sequences of 60 seconds. The decision to use shorter sequences aims to simplify the coding of the model and to reduce training time. Nonetheless, turn-taking is locally managed on a turn-by-turn basis (Sacks et al., 1974). With turns of approximately 3.5 seconds in length on average in the dataset, it is reasonable to assume that the information necessary for predicting floor control should be available in the last

[4] Implementation is based on `https://github.com/patriceguyot/Yin` and is available with the analysis code.

[5] This representation of semitones is known as MIDI note. It was chosen out of convenience.

10 seconds of the conversation. The target for training is the value of the voice activity feature in the following 3 seconds (60 frames).

The model's architecture consists of 10 LSTM input nodes, with `tanh` activation function, and *L2* regularization of 0.001. These were followed by one densely connected node with a `sigmoid` activation function. The loss function is mean squared error, the optimiser is an `RMSprop` with a learning rate of 0.01.

The model is trained over 100 epochs. The Truncated Backpropagation Through Time procedure that Skantze employs over sequences of 60 seconds with an input length of 10 seconds is omitted. As explained above, using sequences of 10 seconds as the input instead should produce a similar result. Another deviation from the original model is the batch size for training. A batch size of 512 sequences is used instead of 32. When trying to train the model with batch sizes of 32, 64, 128, and 256 sequences it failed due to the exploding gradient problem.[6]

After training the floor is calculated as suggested by the original paper: The audio data is processed by the two models, one per interactant, to calculate the probability of an interactant to vocalise in the next frame. The interactant with the higher probability is determined to be the floor-holder.

Partial LSTM

This model is identical to the LSTM model above, except for not using the golden voice-activity as a feature. By omitting the ground-truth from the features array it aims to provide a fairer comparison to the FCD algorithm.

Random

The random algorithm is used as a baseline estimate of chance performance. It is a simulation in which floor control alternates between the interactants. The duration of each floor control segment is exponentially distributed with the average duration equal to the average floor control duration found in the dataset. The average value, 3.53 seconds, is calculated from the annotated floor described above, for the entire dataset.

3.4.4 Performance measures

Four measures are applied to test the algorithms' performance on a test set of 8 dialogues. They intend to highlight the different properties of the algorithms. The averages across the dialogues are summarised in table 3.1.

[6]The exploding gradient problem is apparently a known issue with this type of neural networks. It is somewhat documented online: `https://stackoverflow.com/a/37242531/1224456`, `https://github.com/keras-team/keras/issues/2134`.

algorithm	Accuracy	Backchannels	Stability	Lag
FCD	0.857	0.871	0.908	0.422
Optimised FCD	0.874	0.721	0.470	0.105
LSTM	0.888	0.816	0.291	0.163
Partial LSTM	0.871	0.770	0.243	0.169
Random	0.483	0.509	0.961	1.757

Table 3.1: Summary of four performance measures for the floor control detection algorithm, with default and optimised parameters, and the competitor algorithms.

Accuracy

The accuracy measure represent the ratio of frames in which a algorithm agrees with the ground-truth floor described in Section 3.4.2. The annotated floor is compared to the prediction of each algorithm every 10 seconds. This time interval has been chosen to make sure the samples are independent.[7] For each frame without competition for floor nor silence the algorithm's prediction is compared with the ground-truth floor to determine if they agree with each other or not. Averaging across all of these frames per dialogue produces the accuracy value.

Backchannel responses classification

The reported floor-holder is extracted for the starting point of every ground-truth backchannel. In line with the definition of floor-holding, if the backchannel does not belong to the floor-holder predicted by the algorithm it is correctly categorised, and incorrectly categorised otherwise. The ratio of correctly categorised backchannel responses is calculated for each dialogue.

Stability

An algorithm for floor control detection should not report unnecessary changes in floor control. One can say that such algorithm is unstable. To calculate stability the number of floor changes in the ground-truth floor is divided by the number floor changes reported by each algorithm. Values close to 1 indicate that the algorithm reports changes in floor at a similar rate to that of the ground-truth annotations. Smaller values indicate that the algorithm reports too many floor changes (i.e. unstable), and larger values indicate that the algorithm predicts too few floor changes (i.e. unresponsive).

[7] See discussion about the locality of turn-taking in Section 3.4.3.

Lag

Another requirement from these algorithms is to have a minimal lag. When the floor-holder changes, the algorithm should indicate that as quickly as possible. In some cases, when the algorithm predicts changes, this can be instantaneously, or even with "negative lag". These negative lags are ignored in this study. For every floor holding segment reported by an algorithm, if, at the start of the segment, the ground-truth floor agrees with the algorithm, the lag is the time from the beginning of the annotated floor segment to the beginning of the reported segment. If the algorithm's floor control decisions are unstable within a segment of the annotated floor (i.e. reporting multiple floor changes) only the first lag is considered and the rest are ignored. These lag values are averaged across each dialogue.

3.4.5 Optimising the algorithm's parameters

The FCD algorithm relies on two parameters: the filters' cut-off frequency f and the comparison threshold k. The optimal parameters for the best accuracy according to basin-hopping optimisation (Wales and Doye, 1997) on the training set are cut-off frequency $f = 1.789Hz$ and comparison threshold $k = 1.067$. The performance on the four measures, calculated on the test set, is presented in table 3.1.

3.4.6 Estimating the differences between the algorithms

Table 3.1 displays the averages each algorithm achieved on each performance measure. It does not, however, indicate how different the algorithms actually are from each other. A Bayesian approach is chosen to evaluate these differences. For each algorithm and performance measure combination the mode is estimated using the test set as followed.

The accuracy measure indicates the ratio of frames that agree with the ground-truth floor. These frames are modelled as Bernoulli distributed with a parameter from the dialogue mode. The parameter for the dialogue mode is beta distributed with mode sampled from the global mode, and concentration sampled from the global concentration. The prior for the global mode is a uniform distribution from 0 to 1 and the prior for the global concentration is gamma distribution (shifted by two so concentration is always >2) with mode and standard deviation of 20. The backchannels measure is modelled the same way.

As opposed to the first two measurements above, the stability measurement produces only one value per dialogue. This value is modelled as gamma

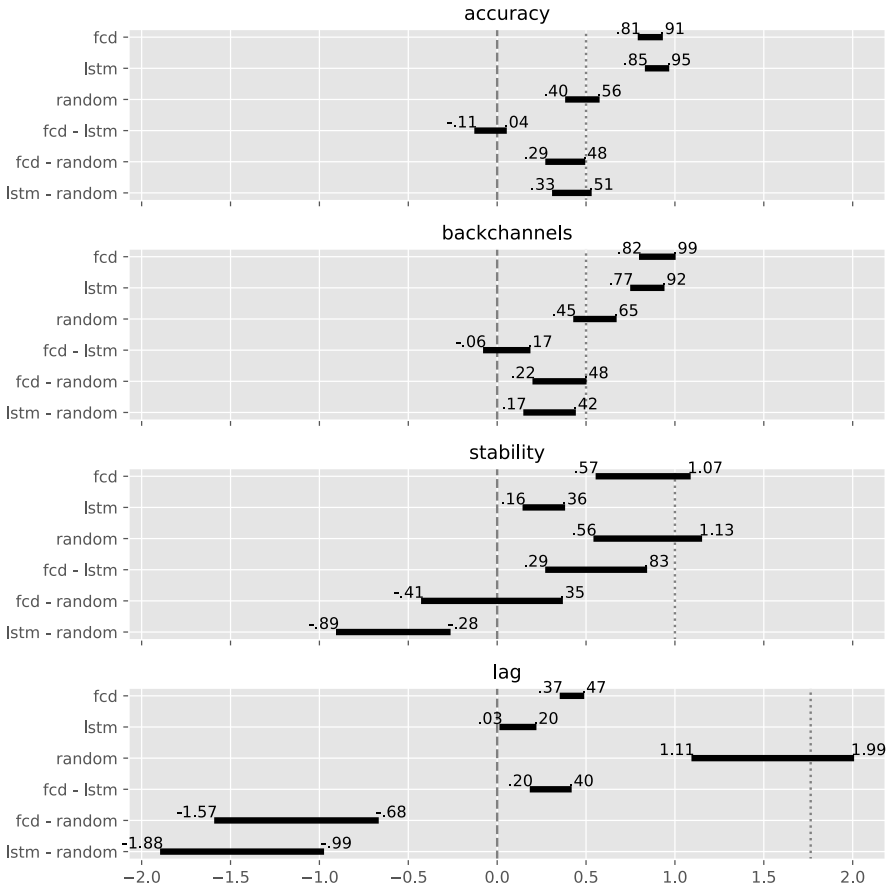


Figure 3.3: 95% highest density intervals for the performance measurements' modes of the FCD, LSTM, and random algorithms, as well as the estimate for the differences between the modes.

distributed with mode and standard deviation sampled from the global mode and global standard deviation. The prior for the global mode and the global standard deviation are Gamma distributions with mean and standard deviation equal to one.

Lastly, each lag is modelled as exponentially distributed with mean taken from the dialogue. Dialogue mean lags are modelled as gamma distributed with mode and standard deviation sampled from the global mode and the global standard deviation. The priors for these are, again, gamma distributions with mean and standard deviation of one.

Figure 3.3 shows the estimates for the algorithm's modes for the performance measurements. These are illustrated in the form of 95% highest density intervals, indicating the values that are the most probable for the modes. Only the FCD with the default parameters, the LSTM, and the random algorithms are presented here for brevity. The difference between the algorithms is considered insignificant if the difference crosses $x = 0$ (indicated by a dashed vertical line).[8] The vertical dotted line marks the expected chance performance.

Apart from the comparison in Figure 3.3, comparing the LSTM algorithm and the partial LSTM model shows that the difference between them, on all performance measurements, are insignificant. Similarly, comparing the FCD with the default parameters to the optimised FCD suggests that the difference in accuracy is insignificant. The lag improves due to the optimisation, while the default parameters perform better for backchannel responses categorisation and stability.

3.5 Examples of floor control detection

Figures 3.4 and 3.5 show the output from the FCD algorithm with the default parameters and the LSTM model, for two short snippets from the dataset. The figures present the audio wave for each interactant at the top. Next they show the annotations, including the ground-truth floor and backchannels, as calculated from the utterances as described in Section 3.4.2. At the bottom there are the outputs from the algorithms. Items coloured in blue represent interactant A, and those coloured in red represent B. These can be audio, utterances, or floor control segments.

The 13 second snippet shown in Figure 3.4 demonstrates some of the main differences between the FCD and the LSTM algorithms. Whereas they both mostly agree with the annotated floor, the LSTM model's instability is shown as it reports many unnecessary floor changes during this snippet. The rela-

[8] Regions of practical equivalence are excluded from the analysis for simplicity. Note, however, that the reported results won't change if these would have been included.

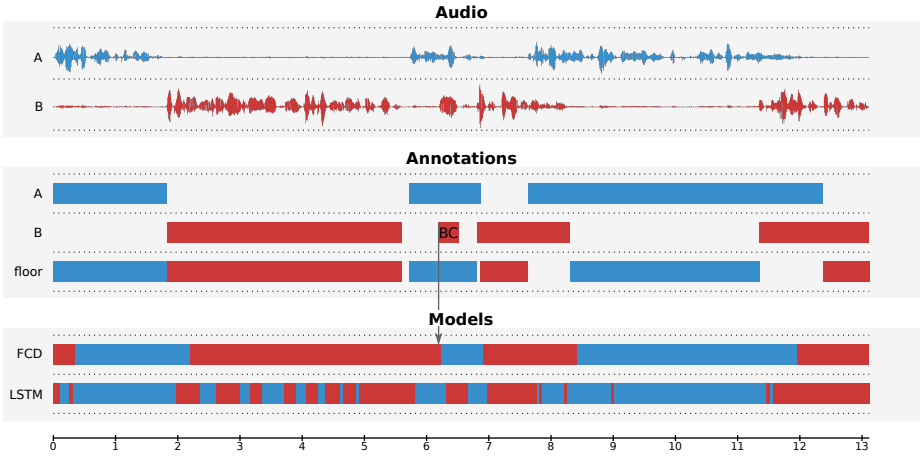


Figure 3.4: Thirteen seconds snippet from a conversation by two interactants, showing the floor control decisions by the FCD and LSTM algorithms compared to the annotations and the audio.

tively longer lag of the FCD algorithm is also presented here. For example, approximately at 2 seconds the floor changes from interactant A to B. It takes the LSTM model less time to report that change than to the FCD algorithm. Lastly, there is a backchannel by interactant B at approximately 6 seconds. On the one hand, the FCD algorithm lags in indicating the floor change to interactant A, thus miscategorises the backchannel. On the other hand, the LSTM categorises the backchannel correctly only to report a wrong floor change immediately after.

Figure 3.5 shows another snippet that demonstrates the behaviour of the FCD and the LSTM algorithms. As demonstrated before, the FCD algorithm has some lag, but is less jittery than the LSTM model. This time there are three backchannel responses by interactant A during interactant B's turn ('ja', 'ja', and 'ja ja'). The FCD algorithm keeps reporting the floor correctly without reporting the backchannel responses as turn changes. Between approximately 5 and 7 seconds it seems that there is a competition for the floor. In this case, the FCD algorithm reports interactant B as the floor holder while the LSTM model predict multiple floor changes. It is hard to argue which output is favorable in this context, and ultimately this should depend on the application.

3.6 Discussion

The accuracy measure represents the ratio of frames in which an algorithm reports the same floor-holder as the one annotated in the dataset. The random algorithm's accuracy is about 0.5. This is expected, and can be explained an-

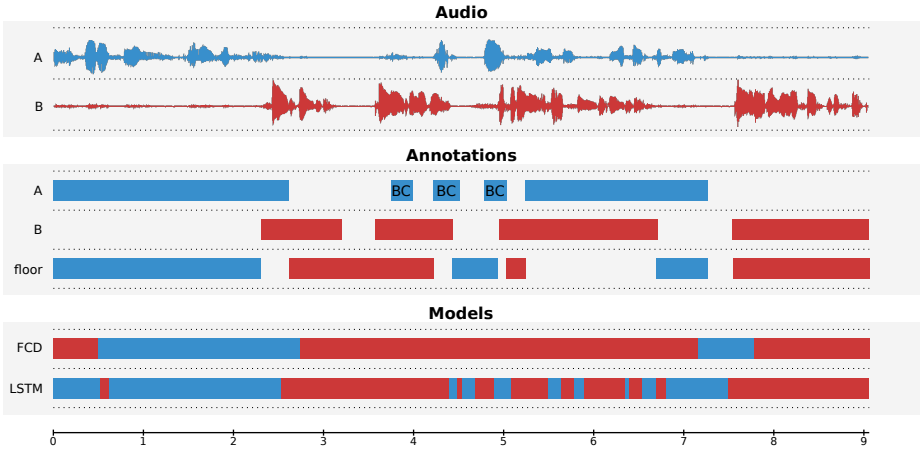


Figure 3.5: Another snippet of a conversation from the dataset. In these 9 seconds there are extended periods of overlap. It shows how the algorithms (FCD and LSTM) cope with a more complex turn taking example.

alytically: a decision by an algorithm, per frame, can either agree or disagree with the annotations. If we assume that each participant holds the floor, on average, half of the time, a random guess will be correct half of the time. The FCD and the LSTM algorithms perform better than random, or chance, in this context, but without significant difference between them.

Backchannel responses are correctly classified if an algorithm reports them as happening when not holding the floor. The random algorithm, in this case, is correct about half of the time. Following the same reasoning from the previous paragraph this performance is expected by chance. As in the accuracy case, the FCD and LSTM algorithms' backchannel responses categorisation is better than chance, with insignificant difference between them. It is important to note that the LSTM model is not designed for this task, and still performs it surprisingly well. The FCD algorithm, on the contrary, is designed with this task in mind. The low-pass filters essentially slow down the changes in loudness (the RMS values). Therefore, on one hand, short utterances are unable to gain a significant increase in level past the filters, and on the other the level at the end of a long utterance drops quite slowly. As a result, the comparison of loudness levels past the filters usually reports the interactant that had the longer utterance as the floor-holder for a while.

The working definition for backchannel responses includes short utterances with the word 'ja' (loosely translated to 'yes'/'yeah'). This was suggested by a native German speaker after investigating one dialogue from the dataset, as discussed earlier in Section 3.4.2. Such utterances might also occur as a positive answer to a question, which usually won't be considered as backchannel

responses. This is a limitation of the automatic ground-truth extraction method employed here.

The stability measure reflects the effect of the filter and the comparison threshold. Here, the FCD algorithm with the default parameters performs better than the LSTM model and the optimised FCD algorithm. Note, however, that it is not better than the random algorithm. The latter takes the average floor duration into account, and shifts floor between interactants randomly while maintaining this average. It is expected to produce the same number of floor changes as the annotations and therefore have a stability value close to 1. The filters, as highlighted earlier, smooth out and slow down the changes in the power of the audio. The comparison with the threshold makes sure that close values do not trigger multiple erroneous floor changes, which would reduce the stability. No similar mechanism is presented for the LSTM model. As a consequence it displays considerable jitter. It applies a relatively simple comparisons between values that are produced independently for each interactant. Similar techniques to the ones integrated into the FCD algorithm, namely, the filter and the comparison with a threshold, could have been integrated into the LSTM comparison to increase stability. Testing this, however, is outside the scope of this study that aims to compare the FCD algorithm to an off-the-shelf state-of-the-art approach.

The lag measures the time it takes from the moment a floor change occurs till the moment an algorithm first reports the change. Follow a similar reasoning as before, the random algorithm's lag is expected to be close to half the average turn duration, at 1.76 seconds. All algorithms performed much better than the random algorithm in that regard, with the optimised FCD algorithm and the LSTM models performing best, followed by the FCD with the default parameters. One of the main goals of the LSTM model is to predict turn changes, not only to react to them. Therefore, it is not surprising that it achieved a relatively low lag. The FCD algorithm was originally designed with some lag, coming from the filters, as a way to correctly classify backchannel responses and increase stability. Therefore, the lag is expected. Optimising the FCD algorithm increased the cut-off frequency f from 0.35Hz to above 1.5Hz, reducing the effect of the low-pass filters, and decreasing the lag on the expense of stability.

The optimisation of the FCD algorithm for highest accuracy did not improve accuracy significantly. This is perhaps because the algorithm with the hand-picked parameters managed to achieve the almost maximal accuracy performance possible with this algorithm design. The optimisation, however, improved lag on the expense of stability and backchannel responses categorisation. Note that there is no one ideal FCD algorithm. The design and optimisation should depend on the intended application. Driving non-verbal behaviour for an

embodied conversational agent (ECA), for example, might require minimal lag, whereas automatic dataset annotation would benefit from maximum accuracy.

The LSTM model achieves its results by exploiting features that the other models do not employ. First, the ground-truth floor annotation of the previous frame is one of its inputs. The development of the partial LSTM model aims to eliminate that advantage. This model is identical to the LSTM model but with the ground-truth floor annotation omitted from its features list. The performance of the partial LSTM model found to be very close to the performance of the LSTM model. Second, the LSTM is not strictly a real-time model as many of the features are z-normalised. Implementing an incremental variant of z-normalisation is possible but also might affect the results. Testing this is outside of the scope for this research.

A model for turn-taking should perform consistently across multiple kinds of conversations, tasks, and contexts. Machine learning (ML) models like the LSTM are sensitive to the characteristics of the training dataset. As for any black-box deep learning model, it is unclear *a priori* how well the trained model will perform in a different context. The FCD algorithm, on the other hand, uses signal processing with two main parameters designed around the way turns are organised in conversation. In principle, it should perform well on other datasets. An interesting example is the case of multi-party conversation. With some modifications[9] the FCD algorithm should perform well in this case, whereas the LSTM model would probably require both significant modifications and retraining. Similarly, the FCD algorithm is anticipated to be more robust to audio bleed between the interactants' microphones than the competitor algorithms. Validating these assumptions remains outside the scope of this work.

Whereas the simplicity of the FCD algorithm is its main selling point, it has some clear weaknesses. First, it cannot predict floor changes, but react to them. The LSTM model predicts the probability that an interactant will speak, and as demonstrated here, can also predict floor changes. Predictive models can be used in multiple applications that are outside the capacity of the FCD algorithm like the development of spoken dialogue systems.

Second, the FCD algorithm operates with some lag when achieving higher stability. A real-time system that relies on the FCD algorithm with the proposed default parameters would report floor changes on average 400ms after they occur. Depending on the context, this may or may not impact task success. For offline use, like in the case of automatic analysis for corpus studies, shifting the

[9]The main and possibly only modification for adapting the FCD algorithm for multi-party conversation is to change the comparison component to check if the maximal RMS divided by the next maximal RMS value (instead of the minimal RMS value) would be larger than the threshold. Note that for dialogues this adaptation results in an equivalent algorithm, because the next maximal value and the minimal value are the same when there are only 2 values.

reported floor changes backwards in time might produce more accurate results. This idea, however, requires further investigation to assess its accuracy.

Third, one of the downsides of the way the FCD algorithm segments the conversation is that it provides no details about overlaps or pauses. If there is an overlap of speech the algorithm reports the interactant that speaks louder as the floor-holder. This simplistic view misses much of the structure that can be found in overlapping speech (Schegloff, 2000). It also provides very little information about pauses. Note that this is less of a result of the algorithm itself and more a consequence of the way floor control is defined in this work. On the other hand, together with utterance annotations, like the ones shown in the examples, or VAD output, the algorithm can provide interesting insights about pauses. For example, a pause can be associated with a floor-holder, showing who is expected to speak.

Lastly, conversation contains much more than just speech. Laughter, crying, exaggerated aspiration and so on are common features accompanying talk (Hepburn and Bolden, 2013). The FCD algorithm was not designed with these in mind, and is expected to make mistakes when analysing, for example, a segment of laughter. More about that in Chapter 6.

As a general outcome of this work, it demonstrates how simple signal processing techniques and design driven by knowledge of real-time dialogue, can compete with much more sophisticated ML systems in the analysis of behavioural data.

Chapter 4

Manipulating head nods in virtual reality

This chapter proposes a virtual reality (VR) system for studying non-verbal behaviours. The system enables the development of what could be described as partial Turing tests: models of non-verbal behaviour are implanted into the virtual avatars of the participants and override parts of their behaviours on demand. A gamified study design encourages the participants to rate these models. The system is used in a study to compare three common theories that explain head nods in free conversation: mimicry, backchannel responses, and listeners' responses to speakers' trouble.

Non-verbal cues are an important aspect of communication, both in face-to-face conversations and in upcoming telecommunication technologies such as social VR. Generative models can support the recreation of non-verbal cues in virtual environments, and the development of embodied conversational agents (ECAs) that can naturally converse with people. Various theories aim to explain non-verbal behaviours in conversation. For example, mimicry theories argue that people unconsciously and automatically mimic each other in conversation. Another branch of research shows that non-verbal responses such as head nods are timed to satisfy the speaker, signalling acknowledgement or understanding by the listener in crucial moments in the conversation. These theories can often provide alternative explanations for the same observed behaviour, but generally predict movements that fundamentally differ from one another. Furthermore, comparisons between this type of theories are rare. In this chapter three theories that aim to explain head nods are compared. These are nods as mimicry, backchannel responses, and listeners' responses to speakers' trouble. Extensive

overview of each of these theories is available in Sections 2.3–2.5.

Research of non-verbal behaviour often, but not always, analyses free conversations. In this case, the analysis can rely on annotations (Goodwin, 1979; Boholm and Allwood, 2010; de Kok, 2013), objective measures like motion capture data (Hadar et al., 1983a; Wu et al., 2020), or a mixture of both (Healey et al., 2013). The downside of these approaches is that it is often hard to argue for causal relationships without the ability to systematically manipulate non-verbal behaviours (Brady, 2011). As a naive (and rather unrealistic) example, consider study results showing that listeners tend to nod immediately after a head nod by the speaker. This can be seen as mimicry, but it can also be explained by a tendency of speakers to nod when yielding the turn and of listeners to nod when taking the turn. Another approach is to systematically manipulate the non-verbal behaviour of one interactant, and measure the effect on their conversational partner. Considering the example above, adjusting the timing of the speakers' head nods can tell if the listeners' mimic or rather nod at turn transitions. To implement this researchers usually employ trained confederates (Bavelas et al., 2000; Chartrand and Bargh, 1999; Tiedens and Fragale, 2003), or ECAs (Bailenson and Yee, 2005; Gratch et al., 2007). Study designs with confederates can be problematic as their behaviour might be significantly different than naive participants (Kuhlen and Brennan, 2012). In studies with ECAs the conversation is always restricted in one way or another. For example, the ECA can take a prescribed role of a speaker or a listener. This is mainly a technical limitation, as ECAs are not yet able to converse freely with people, neither in robotic form nor in virtual environments. Another way to frame this is that they do not pass a conversational (as opposed to textual) Turing test. This results in questionable ecological validity.

To overcome these challenges this chapter introduces a novel VR method. Its goal is to compare perceived plausibility of head nods generated by models. The models in question, for this chapter, are based on the theories mentioned above. Nevertheless, the method described here can be applied to study a wide range of non-verbal and multimodal behaviours. The models are embedded into a collaborative virtual environment. Users in the environment can give control over their avatar to models that generate non-verbal responses. Doing so is described as "faking attention": the user still attends the conversation, but their body language is overridden by an algorithm. A game-like scenario encourages the users to detect fakers, and by doing so, they effectively provide perceived plausibility ratings for the models. This setting enables direct experimental tests of different models of non-verbal behaviours, implemented as alternative algorithms for controlling the avatars.

4.1 Head nodding models

The models below track aspects of the conversation in real time to generate timed head nod predictions. They are incremental in the sense that only past data is considered, and operates in real time in the sense of not accumulating lag due to processing constraints. The choice of real-time incremental models is a necessity in this chapter, as they are deployed into a system that manipulates non-verbal behaviours on the fly. Nevertheless, they pose some general theoretical advantages. First, when we interact with other people we only have the current information about the world, and we process it incrementally. Incremental models can therefore be similar to cognitive processes, at least in that very specific sense. Second, real-time models can be used in a vast range of applications that non real-time models cannot cope with. One prominent example is the development of ECAs. The models below *fire* when a head nod is expected. For example, a head nodding mimicry model tracks the partner's head movement until a head nod is detected, waits a pre-defined amount of time, and fires a prediction for a head nod.

4.1.1 Head nods detection

Head nods are detected based on the vertical position of a head mounted display. Appendix A lists a pseudo code for the model. To guarantee a constant rate, the vertical head position is up-sampled linearly to 100Hz (i.e. a sample every 10 milliseconds). These samples are then filtered with two second order Butterworth filters: a low-pass at 4Hz followed by an high-pass at 1Hz. The model predicts a head nod if the result is smaller than -4 millimetres. The system won't report another head nod until the movement stabilise (a sample between -2 and 2 millimetres). This technique is based on Healey et al. (2013): "low frequency movements (1Hz and below) and high frequency movements (4Hz and above) were eliminated ... head nods were identified as vertical movements at a speed >0.3 mm/frame". The authors did not specify what type of filters are used. The choice of Butterworth filters is due to their ease of implementation and their appearance in the literature in the context of head movement filtering (Hale et al., 2019). In addition, the above study processes data at 60 frames per second. The threshold of 0.3 millimetres per frame therefore translates to 0.18 millimetres per millisecond, or 1.8 millimetres per sample at the current 100Hz sample rate. The detector was tested manually with this parameter by actively trying to nod or move the head without nodding while wearing the head mounted display. This showed that a value suggested by the literature is too sensitive. It was increased to 4 millimetres per samples, as described above, to make sure detections do not trigger by random head movement.

4.1.2 Backchannels

Ward and Tsukahara (2000) suggest a prosody based backchannel responses prediction model. Their model, summarised in Table 2.2, provides a set of hand-crafted rules that use the pitch of the speaker's voice, and speaking start and stop timing, to predict hearers' responses.

This model is chosen for its simplicity and ease of interpretation. Whereas machine learning (ML) based model (e.g. Morency et al., 2008) clearly outperform it, they are usually harder to interpret. This property of the model is especially important when comparing competing models that represent theoretically different non-verbal behaviours.

The model used here has a slight deviation from the one in the original paper. The original model waits for a region of pitch with less than the 26th-percentile pitch level (rule P1 in Table 2.2). This assumes access to the pitch data of the entire conversation, and therefore implies a non incremental operation. To adapt the algorithm to incremental processing appropriate for real-time applications a rotating 10 seconds buffer is introduced. This buffer always keeps the pitch profile of the last 10 seconds. Percentile calculation is done against this rotating buffer.

The model is implemented in the programming language PureData (Puckette, 1996), and wrapped by a server to simplify integration with various software systems[1]. One instance of the model is run for each participant. It analyses their voice and predicts backchannel responses for their conversational partners. The inputs of the model are the audio stream from the participant microphone, and binary variable indicating if the participant is currently speaking or not.

4.1.3 Mimicry

Whereas many studies of mimicry are normative rather than generative, mimicry research in VR often implement it with 4 seconds lag. This lag was recently questioned by Hale et al. (2019), who suggest a shorter lag, of 600 milliseconds, which is the value used in this study. In other words, a head nod is predicted 600 milliseconds after a head nod is detected for the conversational partner, as described in Section 4.1.1.

Note that the mimicry model here specifically deals with head nods mimicry, and not head movement mimicry as most mimicry studies suggest. There are two reasons for this decision: First, this chapter aim to compare mimicry and backchannel responses as possible explanations for communicative head movement. Therefore, to obtain a meaningful comparison, both models should

[1]The backchannels algorithm and server are available online at `https://github.com/Nagasaki45/backchannels`.

operate on the same physical movement. Second, the literature implies that mimicry is a general phenomenon that should operate on multiple levels. If the perception-behaviour link is a fundamental cognitive mechanism, as Chartrand and Bargh (1999) suggest, it should also operate on head nods. Furthermore, even when head movement mimicry is investigated, scholars often conclude that head nods are mimicked in conversation. For example, in a study that explored head movement mimicry, Hale et al. (2019) concluded that "the cognitive mechanisms generating mimicry of head nods act with a constant lag of around 0.588 msec". In addition, whereas head nods mimicry is usually implied by the literature, some studies explore it directly (Wu et al., 2020).

4.1.4 Disfluency

Speaker's disfluencies are detected using IBM Watson's speech-to-text service[2], and the "deep_disfluency" software library (Hough and Schlangen, 2017). Audio from each participant's microphone is continuously streamed to a disfluency detection server.[3] The server sends the audio to IBM Watson, receives a transcription of this audio, passes it through the disfluencies tagger and sends the tags back to the client. Edit term tags (`<e>`) and repair start tags (`<rpS>`) are interpreted as disfluencies by the client.

The time between the timestamp IBM Watson reports as the end of the word, and the time the message is received back to the client, is on average 1 second. The disfluencies tagger internal latency is on average 0.25 seconds, but with spikes going up to 25 seconds. These are caused by the library update mechanism that sends new predictions for previously reported tags based on new information from the speech-to-text service. Nothing here really tells the time from the actual end of the word to the time the message about it is logged. It is impossible to assess this overall latency without transcribing some audio first.

Note that according to Healey et al. (2013) primary recipients react to speaker's troubles in a 1 to 4 seconds delay, suggesting that the networking latency shouldn't be a major issue. This delay was not implemented into the model.

4.2 System description

The system for this project is inspired by social VR applications. These telecommunication applications allow groups of remote users to interact in the same

[2] `https://www.ibm.com/uk-en/cloud/watson-speech-to-text`
[3] Code available at `https://github.com/Nagasaki45/deep_disfluency_server`.

virtual environment. Uniquely to this system, users can press a button that initiates automatic algorithmic control over their avatar's movements. This behaviour is presented to the users as "faking attention", and the time spent faking is referred to as faking periods (FPs) throughout this chapter. Fakers are muted from the audio chat so they still hear everything but cannot be heard.

For every FP one of the 3 models described in Sections 4.1.2–4.1.4 is picked at random. When a model predicts a head nod, a head nod animation immediately starts. The animation is picked randomly from a set of 3 manually designed animations. This variety of fake head nods aims to make the fake behaviours look less repetitive and therefore more realistic. If a new prediction arrives before the previous animation finishes there is a 100 milliseconds cross-fade to the new animation. This behaviour ensures that the fake head nods are both smooth and representative of the output from the model.

Avatars should continue to produce appropriate non-verbal responses while faking. Their hands, for example, cannot just freeze. A few additional mechanisms provide a baseline behaviour to which the automated head nods are added. A basic approximation of attention in a conversation is to always look at the speaker (Fujie et al., 2009). Therefore, faking avatars slowly rotate towards the speaker. The speaker, in this case, is the participant who has their audio chat's voice activation detection (VAD) component turned active the most recently. To keep the head of the automated avatars moving a small amount of randomness is added to the head looking direction. Instead of looking directly to the speaker, the head slowly rotates towards a random point on a sphere around the speaker head. This point is updated every 1.5 seconds, triggering a new slow rotation each time. Lastly, the system always keeps a 4 seconds buffer of the participants hands movement. When the participants trigger the faked behaviour this recorded movement starts playing in a loop. To ensure smooth movement the buffer is first played backwards from the last recorded sample. When the playback reaches the beginning of the buffer the playback direction flips, and this backwards and forwards playback continues repeatedly. The decision to keep the last 4 seconds aims to give a long enough recording to make the faked behaviour look natural, but in the same time not to include old movement data. If the faker was involved in different type of activity a few seconds ago, like speaking for example, it is better to discard these body movements as they are probably inappropriate in the new context.

The system is implemented using commercial VR hardware (HTC Vive[4]) which combines a head mounted display and two hand-held controllers. These components are tracked in 3-dimensional space, allowing the avatars to mirror the users' physical movement. The microphone and headphones' connection on

[4]`https://www.vive.com/uk/`

Figure 4.1: Example view of the virtual environment and the avatar design.

the headset are used for a voice chat between the users. The state reported by the VAD component of the voice chat is sent to the backchannels model as the speaking state input. This system is developed in Unity3D[5], a game engine commonly used to create VR experiences. Its source code is open and available online.[6]

A scoring mechanism is used to incentivise users to fake and detect fakers. Participants see their own score in a floating message in front of them as shown in Figure 4.1. When they fake attention a 'Snake' game[7] pops up above the score. Collecting a snake's food pellet increases the user's score by five points. Another way to get points is by detecting users as they fake attention. A correct detection is worth one point, but an incorrect one loses a point.

Figure 4.2 shows the user interface for controlling the system. Users fake

[5] https://unity3d.com/

[6] The repository of this project is at https://github.com/Nagasaki45/UnsocialVR/tree/study-2. A video demonstrating the environment can be found at https://youtu.be/0Op1pARFM8I.

[7] https://en.wikipedia.org/wiki/Snake_(video_game_genre)

Figure 4.2: The user interface includes buttons to start and stop faking, accuse others for faking, and control the snake game while faking.

attention by pressing a button on the left hand-held controller with their index finger. While faking, the joystick like button under the left thumb controls the snake game. Detecting fakers is done by looking at them and using the index finger trigger on the right hand-held controller. Note that there is no need to point at users to accuse them for faking, as this "pointing and shooting" gesture might interfere with the social dynamics.

The avatars' design can be seen in Figure 4.1. These are cartoon-ish gender-neutral head and hands figures. When the chat application detects voice activity, the mouth of the avatar is animated to open and close in a steady rate. This compensates for the lack of actual tracking of lip movement and to help users to identify the current speaker, as all of the avatars look the same.

4.3 Pilot studies and motivation

Two pilot studies were conducted before the main study reported here. Discussing them provide more detailed picture of how the participants respond to the study and the virtual environment. In addition, they highlight possible mistakes in the development of the virtual environment and can provide insights for future gamified VR based studies.

The two pilot studies used earlier versions of the system. The virtual setting

was a cocktail party on the beach rather than an office. Participants were standing and walking through the environment as the virtual space was significantly larger. The system relied on the *room scale* feature of the VR headset that allows users to move freely inside a predefined physical boundary constrained mainly by the physical space in the room (Gepp, 2017). To support easier movement in the environment a teleportation feature was added. Participants could press a button to aim and teleport to different spots in the environment. Similar teleportation mechanisms are common in commercial VR games and environments (Boletsis and Cedergren, 2019). The participants' task while faking was to collect floating yellow tokens that appeared around them rather than playing snake. These tokens were visible and collectable only while faking. Collecting these tokens required traversing the environment, usually by a combination of teleportation and some amount of physical movement. Similar to the snake game, collecting a box increased the participant's score by 5. Correct detection of fakers resulted in stealing one point from the faker, while incorrect detection resulted in losing one point that is gained by the mistakenly accused participant. Unlike the main study, in the pilot studies the score was hidden from the participants and there was no feedback for correct or incorrect detection of fakers.

In the first pilot study only one automated behaviour was implemented. The experiment was designed to test whether or not participants notice the difference between real non-verbal behavior and an algorithmic one. The second pilot introduces the comparison between the three models: backchannels, mimicry (with 4 seconds delay), and disfluency. Four groups of three participants participated in the first pilot study, and five groups in the second pilot study.

The two pilot studies indicated that the teleportation feature is problematic in the context of the study. Using it frequently disrupt the conversation and triggered many detections. It is assumed that participants found it hard to distinguish between faking and abruptly teleporting away from the conversation. In addition, because the avatars all look the same, teleportation made it harder to know who is who and required repeated checks by the participants to make sure they know who they are talking with. Lastly, informal analysis of the videos suggests that familiarity with VR and the controllers had significant effect on how well the participants used this feature. Proficient users were faster to get tokens using the teleportation feature and made less teleportation mistakes. For these reasons, the setting for the main study was changed to a sitting office environment, and the faking task was changed to the snake game that has no spatial component.

The pilot studies also indicate that the task of detecting fakers was hard:

most FPs went undetected (more on this in Section 4.5.3), and some participants reported that they preferred to collect tokens to increase their score than trying to detect fakers. This diminished the number of detections, resulting in less data points for comparing the models. Feedback for correct or incorrect detections was introduced to encourage the participants to detect fakers. In addition, the score was made visible for the main study. These changes assumed that clear feedback will help the participants to improve in detecting fakers and incentivise them to detect more often and more accurately to increase their score.

4.4 Methods

4.4.1 Participants

Thirty students from Queen Mary University of London participated in the study. They were recruited using mailing lists, and by approaching friends and classmates directly and asking them to participate. Each participant compensated with £10 for their time. The study was conducted in 10 sessions, with 3 participants in each. Data from 5 of these sessions was discarded because of technical issues. The remaining participants were 7 female and 8 male, age 18 to 33 (mean: 23.3, std 3.8).

4.4.2 Procedure

The participants first filled demographic, familiarity with VR, and social closeness questionnaires (available in appendices C.1, C.2, and C.3). Then, they were introduced to the system and the user interface was explained. They participated in a 10 minutes practice phase, together, in the virtual environment, in which they were encouraged to try out the interface and learn how to operate the system. They had no particular task to complete during this practice phase. After the practice phase they were presented with the hot air balloon task, in which they had to discuss an ethical dilemma and reach a consensus (for further details about the hot air balloon task see Howes et al., 2012). They were instructed to try to get the highest score while still attempting to reach an agreement in the hot air balloon task. The experiment phase ran for 15 minutes, during which the VR view of each participant was recorded using a screen capture software, and the participants were video recorded. After it the participants filled the post-experience questionnaire (available in appendix C.4).

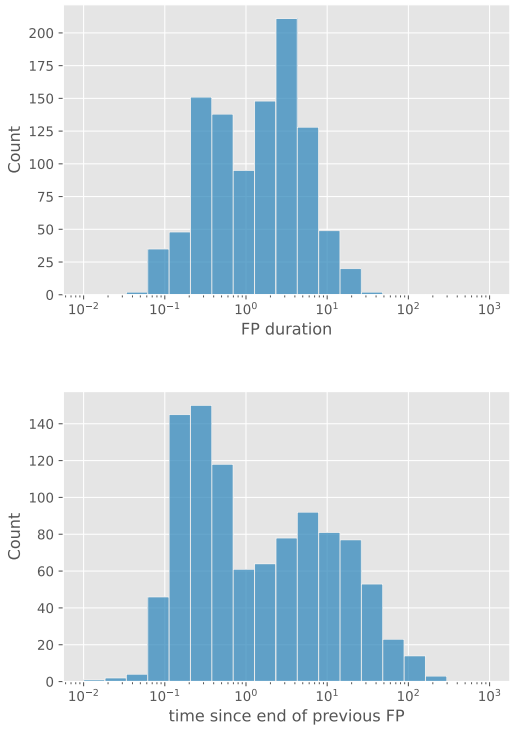


Figure 4.3: Histograms of FPs duration at the top and the time since the end of the previous FP at the bottom.

4.5 Results

As discussed earlier, for each FP a model is picked at random to generate animated head nods for the faker. The FPs characteristics and detection patterns provide insights for how well these models perform.

4.5.1 Faking periods characteristics

Each FP is on average 2.74 seconds, with insignificant differences between the models. The histogram of the FPs durations, shown in the top panel of Figure 4.3, suggests that the duration of the FPs is bi-modal. There are short, below 1 seconds, FPs and long ones, centred around 2–3 seconds in length. The feedback from the participants and manual observation of the screen capture videos suggest that participants start faking and stop immediately as a strategy. When the snake game pops up on their screens they evaluate how easy it would be to collect a food palette. If the food palette is too far away, they often stop and started a new FP hoping that this time the snake's food palette will spawn closer to them. These short FPs are very hard to detect, due to the limited

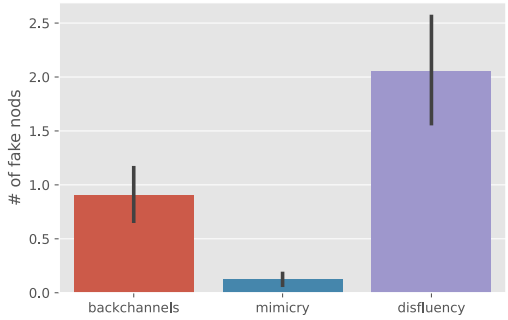


Figure 4.4: The average number of fake head nods generated in each FP by the models.

time the partners have to assess the automated behaviour. Therefore, all FPs that are shorter than 1 second are omitted from the rest of the analysis.

The time between FPs, averaging 8.93 seconds, with insignificant differences between the models, is also bi-modal distributed. As seen at the bottom panel of Figure 4.3, some of them are centred roughly around 200 milliseconds and some around 7 seconds. FPs that start promptly after the previous one ended are probably harder to detect. The real behaviour that appears between the two FPs might confuse the partner and suggest that no faking occur. Therefore, FPs with less than 1 second from the end of the previous one are also omitted from the analysis.

Figure 4.4 shows the average number of fake head nods generated per FP. Considering that FPs by the different models are generally of the same duration, it clearly shows that the models generate head nods in different rates. This is not a problem by itself, as there is no theoretical reason to assume that head nods generated by mimicry should be similar in rate to backchannel responses, for example. Yet, it is important to note the differences as they might affect the ability of the participants to detect FPs. For example, the mimicry model generated almost no head nods.

4.5.2 Detection of fake behaviour

Figure 4.5 shows the percentage of FPs that are detected correctly at least once per model. On average, 25% of the FPs are detected, with insignificant differences between the models. In other words, no model produces nods that are easier nor harder to detect as less-plausible than the others.

Figure 4.6 shows the detection rate (number of detections per second) for each FP per model. The detection rate is on average 0.61 detections per sec-

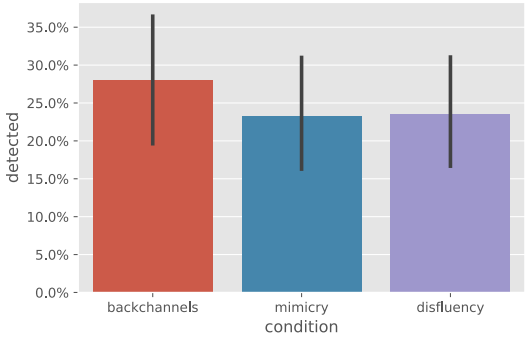


Figure 4.5: Percentage of detected FPs per model.

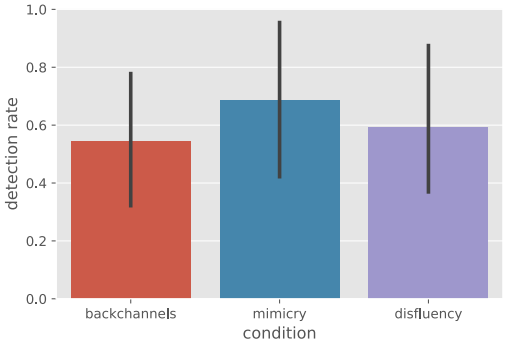


Figure 4.6: Detection rate per model.

	Pilot 1	Pilot 2	Main study
Correct detections	29%	22%	63%
Detected FPs	21%	9%	18%
Time spent faking	22%	14%	21%

Table 4.1: Comparison of results between the pilot studies and the main study.

ond, without significant differences between the models. This value might seem surprisingly high considering the low number of FPs that are detected (25%), and their average duration (2.74). Note, however, that the participants always receive feedback for correct or incorrect detection by sound and visible score. Therefore, when they eventually find a faker, they usually repeatedly trigger detections as quick as possible to increase their score. More on this behaviour later.

4.5.3 Comparisons to the pilot studies

The results of the pilot studies, summarised in Table 4.1, are generally inline with the results of the main study. As in the main study, a low percentage of FPs is detected.[8] Similarly, the second pilot study, which introduces the three models for automating head nods while faking, does not show significant differences in detected FPs and detection rates between the models. The percentage of correct detections, however, increases in the main study compared to the pilot studies. This is a result of providing feedback to the participants for their detections. A common strategy in the main study became to randomly detect fakers until finding one, and then retriggering the detection as quickly as possible on the faker to accumulate points. Needless to say, this strategy is impossible when the score is hidden and no feedback is provided as in the pilot studies.

4.6 Discussion

The system used for this study and the study design demonstrate some of the advantages of VR based studies for social cognition research. Evaluating non-verbal behaviour models is often done by implementing these models into ECAs. Such study designs usually restrict the conversation, for example by assigning the role of the speaker to the participant, and lets the agent produce only listening behaviours (Bailenson and Yee, 2005; Maatman et al., 2005; Gratch et al., 2006, 2007; Hale and Hamilton, 2016). Doing so reduces the ecological validity

[8] In the table all of the FPs are analysed, including those that are shorter than 1 second and those that started with less than 1 second gap since the previous FP. This is for consistency with the data collection and analysis of the pilot studies.

of the results, as they describe a scenario that is different than real-life conversation. The reasons for these restriction are straightforward. ECAs still do not pass a Turing test, hence cannot participant fully in a conversation, and even if they were, restricting the conversation provides a way to test specific components in the ECA implementation (e.g. backchannel responses). Driving the avatar with real human behaviour most of the time, and replacing it temporarily with an automated behaviour, simplifies the agent implementation and allows direct testing of specific ECA components.

Another intended advantage of the gamified VR environment is its ability to directly measure credibility of an automated behaviour, as perceived by participants during a shared task. Many studies that use VR as a method to investigate social interaction rely on self-reported measures. For example, by presenting a mimicking or not mimicking agent to participants who are later asked to rate the agent on multiple forced-choice questions (Bailenson and Yee, 2005; Gratch et al., 2006, 2007; Hale and Hamilton, 2016). Relying on self-reported measures is inherently problematic in social research, as participants often form a post hoc logical explanation for their behaviour (Nisbett and Wilson, 1977).

While this method opens up new possibilities, it also has limitations. Social interaction in VR might be significantly different from face-to-face conversations. This is essentially an empirical question and the answer will change as the capabilities of the technology change. It is important to note, however, that social VR is an increasingly important mode of communication in its own right (Wallis, 2016). Studying communication in social VR might help us understand and build better virtual agents and environments even if the findings cannot seamlessly transfer to the physical world. Another limitation is that the current system uses data from specific hardware with specific capabilities: tracking a head mounted display and two hand-held controllers in 3-dimensional space. This implies that only behaviours that are tracked by the system can be generated by the models and checked for their credibility. For example, facial expressions, eye gaze, fine fingers movement, and torso pose, are not tracked by the system, and cannot be tested. More advanced sensing hardware, however, might improve this in the future.

The results suggest that the participants performance in detecting fakers or differentiate between the model is quite low. Presenting the study as a competitive game contributes to this issue, as participants search for a strategy that maximises their score. In the pilot studies this strategy was to abandon the conversation altogether and search for the yellow tokens. In the main study this strategy is to try to detect others, every few seconds, until a faker is found, and than repeatedly detect them as quickly as possible. Needless to say, neither of these strategies encourage the participants to notice differences in head nods

and provide useful information about the perceived credibility of the models.

Lastly, these studies raise the question of whether or not it is reasonable to expect participants to consciously signal issues with their partners' non-verbal behaviour. Vast amount of research highlight the impact of non-verbal cues, but to my knowledge no studies explore the participants conscious processing of non-verbal communication, especially not when asked about them in real-time during a conversation.

Chapter 5

Comparing models of head nods in conversation

This chapter analyses the predictions from the three theories from Chapter 4, namely: mimicry, backchannel responses, and listeners' responses to speakers' trouble. In two studies, one in face-to-face settings and one in virtual reality (VR), pairs of participants freely converse for 15 minutes while fitted with a motion capture system. Head nods are automatically extracted from the data, and compared to predictions generated by models based on the three theories. The goal of the comparison is to discuss the characteristics of the predictions each theory generates, to understand the correlations between the predictions that originate from different theoretical streams, and to assess how well, overall, the theories account for the observed head nods.

Chapter 4 aims to compare three theories that explain head nods in conversation: mimicry, backchannel responses, and responses to speakers' trouble. It does so using a novel VR based method. In general, the results do not show differences between the theories in terms of their plausibility as generators of head nods. As discussed in length in Section 4.6, there are two main reasons for this. First, the gamified environment did not encourage the participants to notice differences between the theories. Second, consciously indicating if non-verbal behaviour is plausible, in real time, is not a straightforward task.

The current chapter compares the same theories again, but with simplified methods. In the two studies reported below pairs of participants engage in free conversation while fitted with a movement caption system. Their head movement is analysed for head nods and these are compared to the predicted

behaviour according to the three theories above. This way, the predictions can be assessed by their precision (the ratio of predictions that match an actual head nod) and recall (the ratio of head nods that are predicted). The first study takes place in a collaborative VR environment, with the two participants physically located in different rooms. The second study is a face-to-face implementation of the first study. Its goal is to show that the findings are not specific to VR.

In addition, whereas two of the theories in question (backchannel responses and responses to speaker's troubles) are specific to listeners, mimicry theory is agnostic to conversational roles. In other words, the mimicry literature in general ignore speaker-hearer roles: whereas some studies show that the hearer mimics the speaker others present the opposite. This research aims to shed a light on the differences between the theories in respect to speaker-hearer roles. More specifically, it provides ways to assess how well these theories explain speakers versus hearers head nods.

The results of the face-to-face study are inline with the results from the VR study. Therefore, for brevity, and unless noted otherwise, the reported methods and results are for the face-to-face study. Where relevant, values related to the VR study are reported inside parentheses in italics.

5.1 Methods

5.1.1 Participants

Thirteen (*6 in the VR study*) pairs of native English speakers that knew each other in advance participated in the study. Fourteen (*7*) participants identified as female, and 12 (*5*) as male, of 18–26 (*24–56*) years old (mean: 20.8 (*30.8*), std: 1.9 (*9.2*)). Most of the participants were undergraduate and master students in STEM, who were recruited through university mailing lists. Each of them received £10 compensation for their participation.

5.1.2 Apparatus

Two participants at a time participated in the study. In the face-to-face study they were seated in the same room, two metres apart, facing each other. Each participant was recorded by a dash microphone and a video camera that is placed next to their conversational partner. This setup is shown in Figure 5.1. In the VR study they were seated in different rooms and joined a collaborative VR environment that shares the same environment and avatar design, as well as voice chat, as the system described in Section 4.2.

To track head movement in the face-to-face study the participants wear

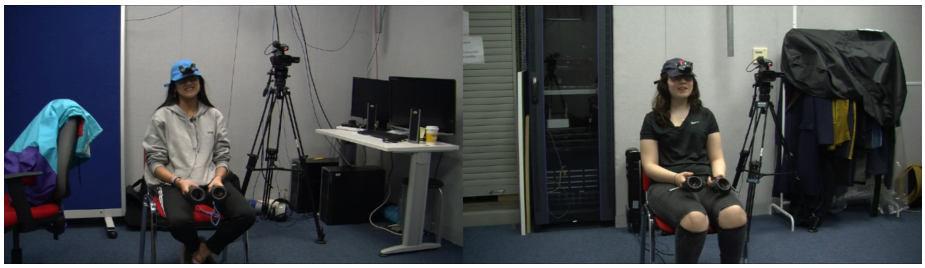


Figure 5.1: A snapshot from a conversation session from the face-to-face dataset.

baseball caps fitted with HTC Vive trackers[1] on their visors. They also held HTC Vive handheld controllers[2] that tracked their hands movement. Usually, the HTC Vive controllers and trackers work alongside the HTC Vive headset. Because the headsets were not necessary for this study they were set aside at a table near the participants and their movement data wasn't recorded. In both studies the logged movement data includes the head and hands position (Cartesian X, Y, and Z values) sampled at the frame rate of the motion data capture application, which varies between 60Hz and 90Hz.[3]

Another difference between the face-to-face and the VR studies is the processing of speech audio. Recording the participants' voices in the same room in the face-to-face study introduced a significant audio bleed. That is, the microphone of one participant recorded a significant portion of their partner's voice. Therefore, and because of the models' dependency on audio, the audio from the microphones is first processed by the iZotope RX 6 De-bleed[4] software to reduce the bleed as much as possible. The reduction strength was set to maximum and the artefact smoothing to 0. No other settings were tested. The reduction in bleed is not tested exhaustively, but a few measurements suggest that after processing the bleed is reduced by up to 40dB. This version of the audio files is used instead of the original audio files for the floor control detection (FCD) algorithm (described next), and the backchannels and disfluency models.

Predictions for the backchannels, mimicry, and disfluency models described in Section 4.1 are calculated per participant. For the backchannels model, the existence of speech in the audio data is detected by the voice activation detection (VAD) application py-webrtcvad[5]. The library is set to process the audio in 30 milliseconds buffers with aggressiveness set to 2. In addition to the mimicry with 600 milliseconds lag, the predictions of a mimicry model with 4 seconds

[1] `https://www.vive.com/us/vive-tracker/`

[2] `https://www.vive.com/us/accessory/controller/`

[3] The application for running this study is available online at `https://github.com/Nagasaki45/F2F-study/tree/study-4`.

[4] `https://www.izotope.com/en/products/repair-and-edit/rx/features-and-comparison/de-bleed.html`

[5] `https://github.com/wiseman/py-webrtcvad`

lag are also calculated. The first is reported below as `mimicry600`, while the second as `mimicry4000`.

5.1.3 Ground-truth

In both studies the FCD algorithm described in Section 3.3 processed the audio from the participants' microphones to determine the ground-truth floor holder at any given moment. The head height of the participants, as captured by the tracker in the face-to-face study or by the headset in the VR study is used to detect head nods as described in Section 4.1.1. These head nods are considered ground-truth for the purposes of this chapter.

5.1.4 Procedure

First, the participants filled a demographic questionnaire (see appendix C.1), followed by a questionnaire about their social relationship (see appendix C.3). Then they were introduced to the Dream Apartment design task (described in detail in Hough et al., 2016), in which they are asked to discuss the design of an apartment for them to share. The exact instructions given to the participants can be found in appendix B. The participants were then fitted with the motion capture system, discussed the task for 15 minutes, until the experimenter asked them to stop.

5.2 Results

5.2.1 Head nods frequency

First, head nods frequency is calculated separately for floor holders and non floor holders. For each participant, the number of head nods while holding the floor is divided by the total time they held the floor to find the nodding frequency. Nodding frequency for non floor holders is calculated in a similar fashion. Comparing these values suggest that floor holders nod more frequently (mean=0.27Hz, std=0.14) than non floor holders (mean=0.19Hz, std=0.09; $t(25) = 3.96$, $p < 0.001$). This is inline with the results from the VR study ($p < 0.05$).

5.2.2 Overlap between models

Here the question of whether the models actually differ from each other is addressed. Whereas they are driven by different theories, there is no reason to assume that the models produce different predictions.

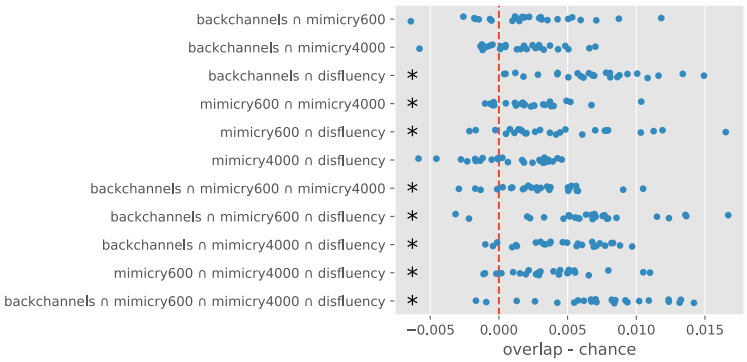


Figure 5.2: The difference between overlap and overlap expected by chance for combinations of models, per participant. Model combinations with overlap significantly higher than chance are marked with an asterisk.

To test that, the overlap for every combination of 2, 3, or all 4 models is calculated. A window of 400 milliseconds is defined around each prediction. The overlap for a combination of models is defined as the percentage of time for which the windows intersect, plus the percentage of time for which no window is reported (i.e. a logical XNOR operation on the windows). Let us consider an example with two models and a total duration of one second. If model A predicts a head nod at 0.2 seconds and model B predicts a head nod at 0.4 seconds the 400 milliseconds windows intersect on the interval 0.2–0.4 seconds and no window is reported for the interval 0.6–1 seconds. Therefore, they agreed on 60% of the time, and this is their overlap.

For each combination of models a chance overlap is also calculated. This is the overlap that is expected from models that output windows with the same summed duration as the windows produced by the actual models, but at random timestamps.

Figure 5.2 shows the difference between the overlap and the overlap expected by chance for all combinations of models. Each dot indicates a participant to visualise the distribution. Most of the combinations of models produce higher overlap than chance. Significantly higher than chance combinations are indicated with an asterisk (assessed by t-tests with Bonferroni correction for multiple tests thus $\alpha < \frac{0.05}{11}$).

From the combinations that has significantly higher overlap than chance in the face-to-face study, the following combinations in the VR study are insignificant: `mimicry600` ∩ `mimicry4000`, `mimicry600` ∩ `disfluency`, and `mimicry600` ∩ `mimicry4000` ∩ `disfluency`.

The significant overlap between the backchannels and the disfluency models, found in both studies, can be interpreted in a few ways. First, these mod-

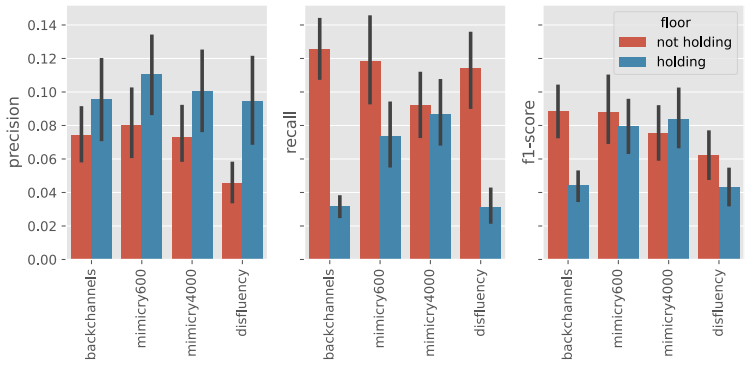


Figure 5.3: Precision, recall, and F1-score for the models' predictions for participants holding or not holding the floor. The error bars indicate a 95% confidence interval of the mean.

els use speech audio as an input and design to predict head nods by listeners. The calculated chance, on the other hand, ignores conversational roles completely. Another possible explanation for the higher than chance overlap is that disfluencies share prosodic characteristics with the backchannels model. This is discussed in depth by the authors that proposed the backchannels model (Ward and Tsukahara, 2000). Furthermore, they claim that disfluencies elicit backchannel responses.

The overlap between the mimicry600 and the disfluency models is also significantly above chance level. This finding is in line with previous studies that found increased gestures' similarity in free conversations during disfluent utterances (Healey et al., 2015). This is, however, only supported by the results of the face-to-face study, as the VR study does not show significantly higher than chance overlap between these models.

No theoretical argument, nor an implementation detail, could explain the higher than chance overlap for the `mimicry600` ∩ `mimicry4000` combination found in the face-to-face study.

Lastly, given the overlaps found between combinations of two models, it is not surprising that combinations of 3 and 4 models also overlap.

5.2.3 Models comparison

Precision, recall, and F1-scores (the harmonic mean of precision and recall) are calculated, per participant, to assess how well the models perform. Because the models' predictions and the detected head nods are timed events the usual precision and recall definitions have been adapted as follows.

- **True positives:** A prediction is correct if a head nod is detected in the

400ms window around it, as suggested by Poppe et al. (2010). To make sure not to inflate the number of true positives no more than 1 true positive per detected head nod is allowed.

- **Precision:** Number of true positives divided by the number of predictions.
- **Recall:** Number of true positives divided by the number of detected nods.

This method is common for backchannel models evaluation (de Kok, 2013), and because the predictions here are similar (temporal point processes) this method should be appropriate.

Figure 5.3 shows the precision, recall, and F1-score of the models while predicting for participants while holding and not holding the floor. The results suggest that there are no major differences between precision values for the different models. On the other hand, all models achieve higher precision when predicting speaker head nods than when predicting hearer head nods. This effect can be explained by the higher frequency of head nods while holding the floor, as discussed before, as this increase the chance of a prediction to match an actual head nod.

The backchannels and disfluency models achieve higher recall for hearers versus speakers. These models are designed for hearers, so it makes sense that they would perform better in that case. In addition, unlike the other models these rely on speech information to operate. The same input is also used to determine the floor holder. By relying on the same input for the models and for analysis the model can achieve arbitrary high recall. For example, predicting head nods in a high rate for hearers and none at all for speakers can produce an almost perfect recall for hearers and zero for speakers. Therefore, conclusions based on recall about the performance of the backchannels and disfluency models in relation to conversational role should be taken cautiously.

The mimicry600 model also has a higher recall for participants not holding the floor compared to floor holders. It might suggest that mimicry of speakers is more common than mimicry of hearers. Another possible explanation is the increased nodding frequency for floor holders discussed earlier. Because the head nods frequency of floor holders is higher than non floor holders, more predictions are generated for non floor holders, and therefore more detected head nods match with a prediction, inflating the recall value. The mimicry4000 model is not expected to be affected by this, as the average turn duration is 3.2 seconds (*4.07 seconds in the VR study*). In other words, the predictions of the mimicry4000 model often happen beyond the turn that triggered them, so the influence of the floor on this model's recall is negligible.

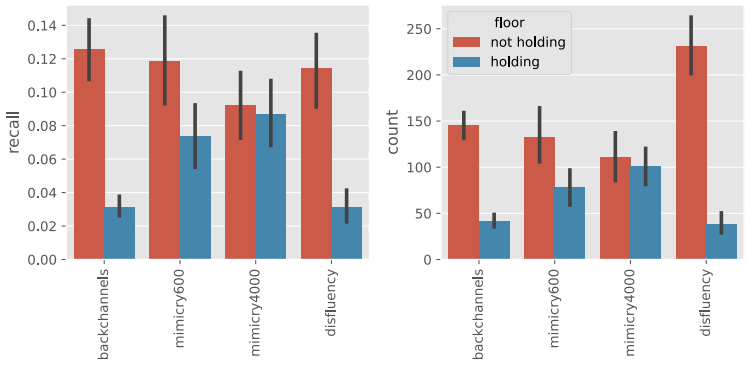


Figure 5.4: Recall performance of each model (left) and number of predictions generated by each model per participant, when holding and not holding the floor (right). The error bars indicate a 95% confidence interval of the mean.

Figure 5.4 provides another explanation for the differences in recall values between the models. The left bar graph shows the recall values of the models, for floor holders and non floor holders (same as the middle panel of Figure 5.3), while the right bar graph shows the count of predictions in each of these conditions. For a fixed number of detected head nods, a higher number of predictions will result in an increased recall value. The similarity between the graphs suggests that the differences between the models in terms of their recall values is a direct result of the number of predictions in each condition. This might disprove the suggestion that different models capture speaker and hearer head nods behaviour better than others, as these differences could be due to prediction frequency alone.

5.2.4 Accounted head nods

This section assesses how many of the detected head nods are accounted by any of the models. A union of all four models is defined by all of the predictions of the models. Considering the definition of recall from the previous section, the recall of this union model indicates how many of the detected head nods are predicted by any of the models.

The average union model's recall across all participants is 0.29 (*0.30 in the VR study*). This low value indicates that the vast majority of the detected head nods are not accounted by any of the models under investigation. The recall for floor holders is 0.20 (*0.22*) and for non floor holders is 0.38 (*0.39*). As two out of the four models, the backchannels and disfluency models, are designed specifically for listeners, the union's recall for non floor holders is expected to be higher. Nevertheless, even for non floor holders most of the detected head

nods in the data are left unexplained.

One possible explanation to this low number of accounted head nods is the 400 milliseconds window used for deciding if a prediction is correct. As discussed earlier, this value is common in research of backchannel responses (Poppe et al., 2010), but it is on the shorter side of the spectrum in the literature (de Kok, 2013, page 70). A longer window should increase this number, though the window size choice should not be decided by this result. Another possibility is that the head nods detection picks movements that are not necessarily head nods. This general issue with the current study is discussed further in the next chapter. Lastly, although the models discussed here cover a few different theories that are extensively studied, it is clear that they do not cover the entire range of head movement and head nods found in free conversations. As shown earlier, speakers nod more than listeners, and no model here is designed to address this specific nodding behaviour. A theory that specifically deals with speakers' head movement and nods might shed a light on the head nods that the models here failed to predict.

5.3 Discussion

This study compares predictions of three head nodding models to automatically detected head nods in free dialogues. It argues that there are significant overlaps between the models, suggesting that the same head nods can often be explained adequately by multiple theories.

The differences between the models and their performance for different conversational roles is not conclusive. All models achieve higher precision when predicting speakers head nods than when predicting hearers head nods. This is expected because speakers nod more and by that increase the chance of a prediction to match a detected head nod. Whereas some theories achieve higher recall for specific conversational roles the difference is usually caused by increased number of predictions generated for these roles.

Most of the observed head nods in the data went undetected by any of the models under investigation. This can be a result of technically insufficient methods of detecting head nods and matching these with predictions. The next chapter addresses this issue. In addition, this research relies on a method to evaluate predictions for timed events that does not produce precision and recall as commonly used, but a slight variations of them. It is used in the literature for similar tasks (Poppe et al., 2010), but no evaluation of this method nor the window choice could be found in the literature. Alternatively, the low number of detected head nods that are accounted by the models might suggest that these theories are not enough to explain head nods by speakers and listeners in free

conversation.

Chapter 6

What is a head nod?

This chapter investigates the definition of a head nod. A naive description of a head nod is a downwards movement of the head, followed by an upwards movement. This is inline with the automatic head nods detector employed throughout this thesis. To challenge this, short video snippets of dialogues containing automatically detected head nods and peaks in head movement are extracted. The video snippets are then manually annotated and analysed to assess agreement between annotators on what is a head nod, and relationships between movement and semantics.

The previous chapter tests how well models that predict head nods fit with automatically detected head nods in conversation. The generally negative results suggest that the mismatch between the predictions and the detected head nods stem from the over simplified head nods detector. One possible conclusion is to question what is a head nod in terms of movement and semantic characteristics.

Head nods are an important communicative cue in conversation. While their function is often discussed (e.g. backchannel responses Yngve, 1970), the description of the movement associated with them is often rather simplistic. Common descriptions found in the literature include: "The main stroke of the vertical head movement is downwards" (de Kok, 2013), "Rotation down-up on pitch axis" (Kousidis et al., 2013), "a forward movement of the head going up and down, which can be multiple" (Allwood and Cerrato, 2003), and "a vertical up-and-down movement of the head rhythmically raised and lowered" (Kapoor and Picard, 2001). When the goal is to automatically process non-verbal communication, for example, for the development of embodied conversational agents (ECAs), a more accurate definition of what constitutes as a head nod is required.

This often results in a set of rules for processing motion capture data to detect head nods (e.g. Cerrato and Svanfeldt, 2006; Healey et al., 2013), or alternatively done by training a machine learning (ML) model from data (Morency et al., 2005).

To complicate the question of what is a head nod further, features related to periodicity of the head movement found to interact with both function and meaning. For example, the number of repetitions in head movement change based on conversational role, and interact with other modalities (e.g. gaze) Poggi et al. (2010); Boholm and Allwood (2010). In addition, up and down head movement happens in a variety of behaviours, not just nodding. These include speech Hadar et al. (1983b) and laughter Griffin et al. (2013) among others. Understanding what behaviours have similar movement properties to nodding can help us understand it better.

Both the simplistic definitions and the more sophisticated head nod detection approaches assume that head nods are well defined, distinct phenomena, that can be found by analysing head motion. This chapter questions that argument. It checks to what degree people agree whether a head movement is a nod. Then, it explores whether a head nod is a description of movement or of contextual behaviour mixing movement and semantics, and investigates the behaviours that share similar movement or semantic characteristics but are otherwise not considered as head nods.

6.1 Methods

This study builds upon the face-to-face portion of the dataset collected and described in Chapter 5.

6.1.1 Points of interest

The analysis here uses a concept of points of interest (POIs). These are timestamps in the conversation that are extracted automatically in two methods, per participant. The first POI type is automatically detected head nods. These are extracted as described in Section 4.1.1.

The second POI type is peaks in head movement. These are calculated from the motion captured head position data. The velocity of the head is calculated for each sample (distance between adjacent samples divided by the sample rate). Plotting the velocity over time while inspecting the videos showed a few outliers per participant. These are sudden jumps in velocity, usually for only one sample. Possible explanations for these are tracking issues of the motion capture system or sample rate inconsistencies. An autoregressive model

	# overlapping snippets	Average distance (seconds)
All snippets	36%	1.09
Detected head nods	20%	1.71
Peaks in head movement	3%	4.94

Table 6.1: Overlap between video snippets for annotations.

is employed as an outliers detector. For every sample of velocity data the model receives the previous 20 samples to predict a value. If the absolute difference between the predicted and the actual value is greater than 0.5 metres per second the actual value is replaced by the previous sample. A simplified method of applying a hard threshold to the velocity data was also tested, but, depending on the threshold value, it either reports too many or does not detect enough outliers. After removing outliers, local peaks in velocity data with minimum of 2 seconds between peaks are extracted, these are the peaks in head movement POIs.

6.1.2 Video snippets

Detected head nods and peaks in head movement are found for each participant. These POIs are marked as belonging to a speaker or a listener based on the floor control detection (FCD) algorithm from Chapter 3. For each of these sets (e.g. detected head nods by participant 1 while speaking) the timestamps of the 5 POIs with the highest head velocity are extracted. In total these are 520 timestamps (26 participants times 2 POI types times 2 floor control states times 5 top velocity instances). For each of these timestamps a 6 seconds video snippet, starting 3 seconds before the timestamp, is rendered. These contain audio from the participants' microphones. Figure 5.1 shows a snapshot from such a snippet. The choice of 6 seconds is based on information from the FCD algorithm that the average turn duration for this dataset is 4.35 seconds. Combined with the argument that turns in conversations are locally managed (Sacks et al., 1974), 6 seconds should provide enough context to annotate the videos.

Table 6.1 summarises the overlap between the video snippets for annotation. The overlap indicates the percentage of snippets that overlap with a previous snippet, while distance is the time between their mid-point timestamps. For the entire dataset, 37% of the video snippets partially overlap with a previous snippet. The average distance for these (not counting snippets that do not overlap) is 1.09 seconds. When analysed separately, the detected head nods and the peaks in head movement snippets overlapped less, and with larger average distances. The head nods detector aims to support the detection of multiple

Attribute	Possible values
Semantic	
Direction	give, elicit
Acceptance	accept, reject
Semiotic function	deictic, non-deictic, iconic, symbolic
Movement	
Body part	head, torso
Sagittal	*binary*
Repetition	partial, single, multiple
Binary	
Head nod	*binary*
Laughter	*binary*
Face touching	*binary*
Distraction	*binary*
Technical issue	*binary*

Table 6.2: Attributes and possible values suggested by the annotation scheme.

repeating head nods, firing once at each downwards stroke. Therefore, overlapping detected head nods snippets are expected. The peaks in head movement are limited to reporting peaks with at least 2 seconds between them, so they overlap less. It seems that most of the overlap in the dataset is caused by both algorithms firing at similar times upon high intensity motion.

6.1.3 Annotation scheme

The annotation scheme used here aims to explore relationships between head movement and their meaning. The manual for annotators is available in appendix D.

The annotation scheme defines a list of attributes and possible values that are filled for each video snippet, considering the middle of the video snippet (time-wise) as the movement to annotate. There are three semantic attributes, broadly copied from the MUMIN annotation scheme (Allwood et al., 2007). These are `Direction` with possible values of `give` or `elicit`, `Acceptance` with possible values of `accept` or `reject`, and `Semiotic function` with possible values of `deictic`, `non-deictic`, `iconic`, and `symbolic` (consult Allwood et al., 2007, for details). These attributes can be filled in or left empty if they do not describe the motion well. Allwood's feedback annotations are especially interesting in the context of head movement as the literature claims that head movement takes significant roles in signalling understanding and/or attention (e.g. backchannel responses). The differentiation between contact, perception, and understanding, and contact and perception (without signalling understanding) is omitted because it seems ambiguous in the context of head movement. The

emotion or attitude attribute is also omitted as it does not aim to explain relationships between form and function. The inclusion of the `Semiotic function` attribute aims to capture the semantics of speakers' head movement. Whereas this attribute is originally designed for hand gestures, all possible values can be found in the dataset. Iconic head gesture often appear when participants describe moving something into something else. For example, by saying "put into the cupboard" while tilting their head and moving it to the side as pushing something with it. This head movement is accompanied by sideways movement of the hands. It is important to note that the annotation scheme asks for the semiotic function of the movement that coincide with the head or torso movement, so it can describe a movement of both hands and head. Symbolic head movements are more rare, but there are some in the dataset. For example, one participant does a gesture of counting money with their hands, and articulate every note they go through with their head.

Another set of attributes describes the physical properties of the movement. The `Body part` is filled in with possible values of `head` or `torso`, `Sagittal` with `yes` or `no`, and `Repetitions` with `partial`, `single`, or `multiple`. If one of these attributes is filled in, all three are required. These attributes and possible values aim to capture a rich set of possible movements while still remaining relatively simple to annotate and to interpret. They are based on piloting a few versions of annotation schemes and finding commonalities between existing annotation schemes. Early experimentation with the dataset suggested that differentiating between head movement and torso movement with only one head tracker is tricky (Chen et al., 2015), and head movements are expected to be very different from torso movements both in terms of form, and of semantic meaning.

Coming from anatomy, the sagittal plane is perpendicular to the ground, separating the body left from right. The `Sagittal` binary attribute splits movement into two broad categories: those within the sagittal plane, and those outside it, as shown in Figure D.1. This plane is often discussed alongside the coronal (dividing the body front to back) and transverse (upper and lower) planes to describe the body and orient anatomical diagrams. There are three reasons for describing a movement within or outside the sagittal plane, and not as belonging to one of the three anatomical planes. First, describing movement along the shared axis of the three planes is ambiguous. For example, a forward movement can be described as a movement along the sagittal plane, but also as a movement along the transverse plane. Second, the three planes cannot describe rotations whereas sagittal and non-sagittal movements can be defined for both displacements and rotations. Lastly, the two most basic classes of head movement found in all annotation schemes are movement on the sagittal plane (e.g. nods and jerks) and outside the sagittal plane (e.g. tilts and shakes). Defining

movements in terms of within or outside the sagittal plan is therefore useful to capture this, and alongside the other movement attributes it can help differentiate multiple types of movements in an easy to annotate and easy to interpret way.

The third movement attribute is `Repetition`. The `partial` value captures movements that start and finish at different positions or rotations without any repetition. These are often associated with posture shifts when associated with torso movement, or head tilts when associated with head movement. Another value is `single` for cyclic motions that finish at the same position or rotation. Lastly, the value `multiple` captures repeated cyclic movements. It is introduced to search for different meanings of single versus repeated head nods, similarly to Boholm and Allwood (2010).

The next two attributes, `Head nod` and `Laughter`, are filled with `true`, or `false`. Leaving these empty is not allowed. Annotated head nods are included to enable comparison to the detected head nods, and to assess agreement on head nods. Early iterations of the annotation scheme showed that many video snippets contain laughter as the source of the movement, hence its inclusion.

Lastly, following the pilots the `Face touching`, `Distraction`, and `Technical issue` binary attributes are collected to drop problematic video snippets from analysis. Participants sometimes touch their face and the hat, and hence move the tracker around without moving their head. There are a few instances of participants rotating towards the experimenter, towards a window when someone passes in the corridor, or receiving a phone call (hence taking their phone out of their pocket) in the middle of the experiment. The `Distraction` attribute aims to capture these. Another possible issue is tracking failures (usually of single or just a few frames) that causes a jump in the tracking data and causes both POI algorithms to fire. The `Tracking issue` attribute aims to record these instances.

6.1.4 Subset for inter-annotator agreement

One annotator (the author) annotated all of the 520 video snippets. Two naive annotators annotated a 52 videos subset to assess agreement and validate the annotation scheme. This subset was selected by randomly picking 13 videos of each POI type (detected head nod and peak in head movement) and floor holding (speaker and listener) combination. The random selection excludes one practice video and 4 examples of snippets containing semiotic functions.

In terms of procedure, the naive annotators received the annotation scheme first, the 4 example videos, one practice video, and a practice spreadsheet. The practice spreadsheet, as well as the main spreadsheet for the annotation of the

Pair of annotators **POI type**	$(1,2)$	$(1,3)$	$(2,3)$
Detected head nod	0.521	0.519	0.553
Peak in head movement	0.487	0.530	0.537

Table 6.3: Average Cohen's kappa agreement per pair of annotators. Annotator 1 is the author.

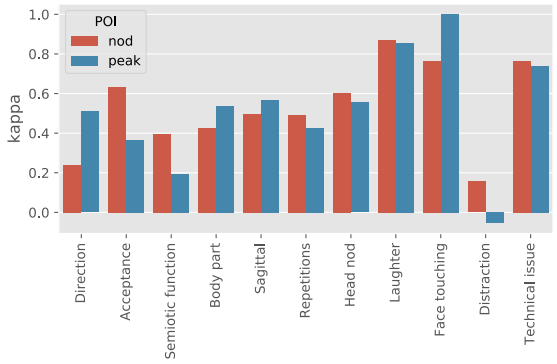


Figure 6.1: Average Cohen's kappa agreement between three pairs of annotators.

52 video snippets, has one column per annotation scheme's attribute. Each row in the spreadsheets is for annotating one video snippet. Learning the annotation scheme, watching the examples, and doing the practice took approximately 45 minutes. This was followed by a short (<15 minutes) video call with the author to make sure the task is clear. Then they received the 52 snippets, each in randomised order, to complete over 2 weeks. The average time to annotate 1 video snippet is approximately 3 minutes. The annotators received £15 per hour for their participation and the durations estimated here are based on the times they reported.

6.2 Results

6.2.1 Inter-annotator agreement on head nods and semantics

The inter-annotator agreement on the 52 video snippets subset is measured in terms of Cohen's kappa agreement coefficient. For every pair of annotators, agreement is calculated separately for each attribute, per POI type. Table 6.3 shows the agreement per pair when averaged across attributes, with the author being annotator 1, and annotators 2 and 3 the naive annotators. No differences between the pairs could be found (post hoc two-way ANOVA with

pair of annotators and POI type as independent variables; $F(pair) = 0.112$; $p > 0.05$). Figure 6.1 shows the inter-annotator agreement averages across pairs of annotators.

In general, the three semantic attributes achieved only fair to moderate agreement (Landis and Koch, 1977), lower than the rest of the attributes in the annotation scheme (except for `Distraction`, more on this later).

Next are the three motion attributes, with moderate agreement (0.4–0.6). High agreement was expected on the `Head nod` attribute, especially for the detected head nods POI type, as these video snippets suppose to include mainly canonical head nods. Although a value of 0.60 for detected head nods is considered as substantial agreement, it is lower than expected here. A lower value here might raises the question of whether or not humans have a widely agreed definition of head nodding behaviour.

The last three attributes are designed for filtering out problematic video snippets. Surprisingly, the agreement on the Distraction attribute is very low. This column is therefore excluded from further analysis or discussion.

The rest of the analysis uses the entire set of 520 video snippets and the annotations by the author.

6.2.2 Excluded data

Out of the 520 annotated video snippets 58 are annotated as containing a technical issue. 55 of these correspond to the participant sitting in the same position in the room, highlighting significant tracking issues in this side of the room.

Another 22 video snippets are annotated as face touching. Whereas face touching can often have communicative meaning, in this dataset it is often related to adjusting the position of the hat with the motion capture tracker, therefore producing significant amount of movement. Also, the participants held handheld trackers, so the expressivity of their hands is assumed to be significantly reduced, suggesting that there is a low chance of them touching their face as a communicative act.

After dropping these the remaining 440 video snippets are used for the analysis.

6.2.3 Annotator agreement with automatic head nod detection

The annotations for head nods provide a way to assess the performance of the head nods detector. Only 16.6% of the video snippets are annotated as head nods. Separated by POI type, 17.8% of the detected head nods and 15.3% of the peaks of head movement video snippets are annotated as head nods.

Annotated head nods **Detected head nods**	`False`	`True`
`False`	15	2
`True`	185	40

Table 6.4: Confusion matrix of automatically detected head nods versus manually annotated head nods.

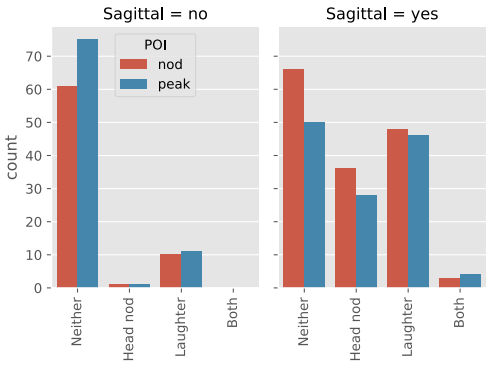


Figure 6.2: Counts of snippets annotated as head nods and laughter for sagittal and non-sagittal movements.

These values are surprisingly low, especially for the detected head nod POI type, suggesting that the head nods detector generally fails to capture head nods successfully.

To assess how many of the annotated head nods are captured by the head nods detector the video snippets in the dataset that do not contain any detected head nod are found. These are 17 videos of peaks of head movements, of which two are annotated as containing head nods. A confusion matrix of detected against annotated head nods is shown in Table 6.4. These values can be converted to precision and recall figures for the head nods detector against human annotations, resulting in 17.8% and 95.2% respectively.

6.2.4 Head nods or laughter

Piloting the annotation scheme suggested that laughter is a significant source of movement in the dataset, hence its inclusion as an attribute. 82.8% of the video snippets that are annotated as laughter are also annotated as sagittal movement, raising the question of movement similarity between laughter and head nods. From the video snippets that are annotated as sagittal, there is no difference between the number of snippets that are annotated as head nods to those annotated as laughter (post hoc Wilcoxon test; $T = 98.0$; $p > 0.05$). This finding suggests that, without additional considerations, sagittal move-

ments are equally associated with laughter as with head nods. Note, however, that although nods are equally likely as sagittal movements, they are rare as non-sagittal movements, whereas laughter can still be non-sagittal, as shown in Figure 6.2.

6.2.5 How behaviours are grouped?

This section searches for meaningful ways of group the different behaviours in the dataset based on similarities between annotations. An analysis of groups of video snippets containing similarly annotated data can reveal relationships between the different attributes within each cluster and especially highlight ties between movement and semantics. For example, clustering together video snippets containing annotated head nods that also give feedback while not holding the floor is an expected outcome.

The kmodes[1] Python library for categorical data clustering groups the video snippets based on the annotations and floor state. Figure 6.3 shows the results obtained by clustering into 4 clusters, per POI type, with one cluster per row and counts of attributes' values per column. The left panel of Figure 6.3a, for example, indicates that most video snippets in cluster 1 are for floor holders (i.e. speaker), while cluster 2 in Figure 6.3b captures more snippets of listeners.

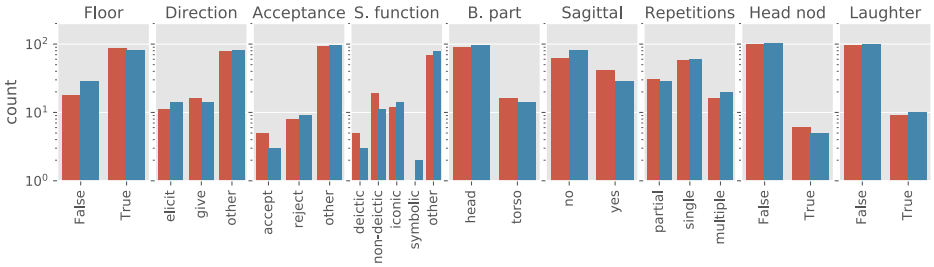
The choice of 4 clusters is arbitrary and is used here to keep the analysis and discussion concise. Values from 3 to 5 clusters have also been tested, so claims that fall short for different number of clusters will be highlighted.

The clusters in Figure 6.3 can be described as follows. Cluster 1 in Figure 6.3a is associated with speakers who mainly move their head and more often in non-sagittal movements of single repetition. These are rarely nods, nor laughter. The rest of the clusters are associated with listeners as indicated by their left-most Floor panels. Cluster 2 in Figure 6.3b captures head nods. In terms of movement, the cluster is dominated by multiple sagittal head movements. It is also the only cluster with `Direction` and `Acceptance` annotated. Cluster 3 in Figure 6.3c captures laughter. With no clear semantics, it is mainly a single torso sagittal movement. The last cluster in Figure 6.3d is for partial torso movements, and can be thought of as posture shifts.

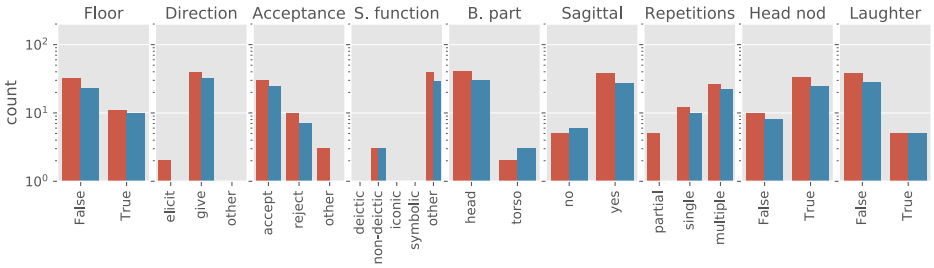
When clustering into 3 clusters one cluster captures a variety of speakers' behaviours. The other two are roughly split by laughter, suggesting that it is the most significant classifier for listeners' intense head movement. While the laughter cluster is similar to the one discussed earlier, the listeners' non-laughter cluster is characterised by repeating head movement, with mixed sagittal values. It gives feedback, that is either accepting or rejecting (not other), and

[1] `https://github.com/nicodv/kmodes`

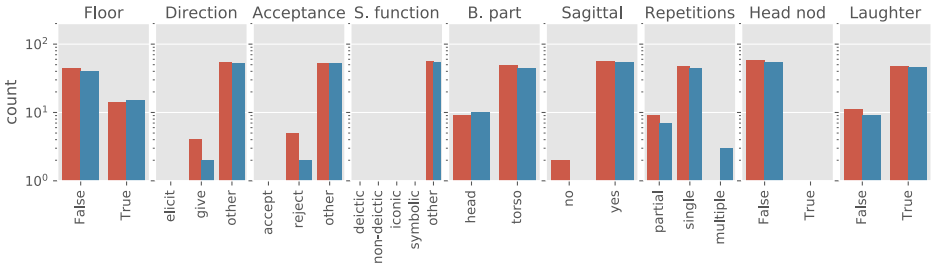
(a) Cluster 1 - speakers.



(b) Cluster 2 - listeners' head nods.



(c) Cluster 3 - listeners' laughter.



(d) Cluster 4 - listeners' posture shifts.

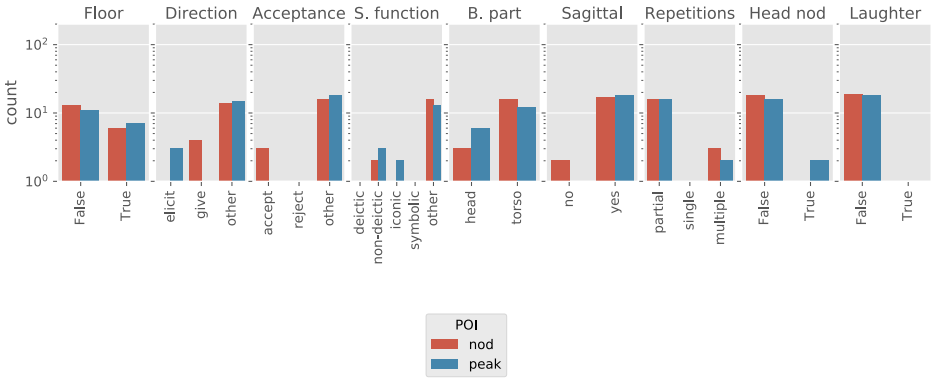


Figure 6.3: Counts of attributes' values after clustering the video snippets by the annotations into 4 clusters, one per row.

contains most of the annotated head nods in the dataset, although the majority of snippets here are still annotated as not a head nod. Generally it seems like a variety of feedback giving, head based, listener responses.

With 5 clusters the 4 clusters from Figure 6.3 re-appear. The new cluster captures a mixture of speakers' and listeners' partial non-sagittal head movement. Similar to posture shifts, these can be thought of as head tilts.

In addition to manipulating the number of clusters, dropping the direction and Semiotic function attributes, which do not achieve moderate inter-annotator agreement (Cohen's kappa <0.4), results in very similar clusters to the original four. This suggests that the clustering is robust enough to produce similar results when the weaker attributes of the annotation scheme are ignored.

6.3 Discussion

This chapter opens with the question of what a head nod is. It attempts to answer this by defining an annotation scheme for head and torso movement and semantics that is informed by the literature and a corpus of video recorded dialogues. Annotations of short video snippets containing dialogues reveal that up and down movements of the head are not necessarily head nods. Sometimes they are, but they are equally likely to correlate with laughter.

The annotations made clear that the automatic head nods detector generally fails to capture the movement people consider head nods. This raises interesting questions regarding the ability to automatically detect head nods, and the nature of head nods in general. To eliminate confusion between head and torso movement, future head nod detection techniques can be improved by motion capturing both and process them to extract head movement relative to the torso. Nevertheless, the relationships between head nods, floor control, and semantics, highlighted by the clustering of annotations, suggest that annotated head nods are contextual and rely on semantic function. This idea that what we perceive as head nods depends on context and is not merely a description of movement is supported by the literature (Nguyen et al., 2012). Therefore, automatic head nod detection should probably rely on semantics as well.

Although the annotation scheme borrows its semantic attributes from the literature, most of them achieved relatively low inter-annotator agreement, questioning their robustness and usefulness. One possible solution for this is to develop head (and possibly torso) specific semantic vocabulary, instead of borrowing from existing annotation schemes that concentrate on hand gestures. This is especially noticeable by examining the annotated semiotic functions. These are predominantly marked as "other", showing that the possible values do not suffice for capturing most of the observed behaviours. In addition, the

clustering of annotations, regardless of the number of clusters, group speakers into one cluster. This suggests that the annotation scheme is not nuanced enough to differentiate and properly describe a variety of head and torso driven speakers' behaviours.

It is important to note that these findings and conclusions are based only on short video snippets containing the most intense movement extracted from longer dialogues (regardless of POI type). One possible side effect of this is the increased focus on laughter in expense of more subtle movements. This effect might have been further exaggerated by using the apartment design task, and recruiting participants who know each other in advance, especially while compared to datasets that use a map task done by strangers (e.g. Paggio and Navarretta, 2010) or formal interviews (e.g. Allwood et al., 2007).

Overall, the findings presented here suggest that research into head nods, and head movement in particular, should go beyond the simplistic view that a head that moves down and up is in fact nodding.

Chapter 7

General discussion

This thesis explores head movement and specifically head nods in free conversation. It begins with questioning which of three common theories for non-verbal behaviour best explains head nods. Chapter 4 discusses a series of experiments that employ a novel virtual reality (VR) method to test this. This method is unique in its ability to precisely swap specific non-verbal behaviours, in real time, with algorithmic alternatives. For example, the head nods of the conversational partners' avatars can be overridden by an algorithm while keeping the rest of their movement properties intact. This can be described as a way to conduct a "partial Turing test": instead of developing an embodied conversational agent (ECA) that can fully participate in free conversation, only the behaviour a researcher is interested in can be implemented to override that non-verbal behaviour of a participant in the conversation. This approach allows researchers to concentrate on the behaviour in question without worrying about the implementation of all the mechanisms that are required to actually pass a full Turing test. Alternatively, and more realistically, it can replace study designs in which the ECA takes a predefined conversational role (e.g. listener) or designs that restrict the conversations in other means. This, in turn, increases the ecological validity of the results. Another goal of this VR method is to allow participates to effectively rate the plausibility of the non-verbal behaviour of their conversational partners in real-time. This approach is unique as most studies in that field that rely on self-reported measures collect these after the experience, usually in the form of questionnaires.

Although this method is motivated by the literature, and while the pilot studies highlighted multiple possible advantages, it uncovered a series of issues that ultimately led to challenge the presuppositions of the original research question. These were initially assumed to be problems with the methodology but eventually raised deeper conceptual issues. The experiments asked the partici-

pants to consciously report the plausibility of the non-verbal behaviour of their conversational partners. On top of that, they were required to do so in real time while actively participating in a conversation, playing a mini-game, and trying to achieve a high score in a somewhat complex gamified study. Whereas many studies highlight our sensitivity to the non-verbal behaviour of our conversational partners there is no claim in the literature that we are able to consciously process these and report when they go wrong.

In retrospect, the idea of conducting these partial Turing tests looks like a strength of this work that should be pursued further. The advantages of VR based methods are often discussed, highlighting their consistency and reproducibility. The ability to manipulate movement of people in conversation in such a granular way should, hopefully, be added to the list of advantages and appear more often in future research. On the other hand, and although interesting, the real-time subjective assessment of the participants' behaviour might not have been the best method to test theories of non-verbal behaviour.

Chapter 5 aims to answer the same research question by eliminating the reliance on the assumption that people can report problems in their partners' non-verbal behaviour. The active head movement manipulation is dropped, and instead of asking the participants to effectively rate head nods, the output of an automatic head nods detector is compared directly to the predictions from the three theories. This also led to the development of the floor control detection (FCD) algorithm discussed in Chapter 3, as it became clear that non-verbal behaviours should be analysed with basic conversational context in mind. This algorithm is employed multiple times throughout this work, facilitating the analysis of head nods.

The results from comparing the theories' predictions to the detected head nods suggest that the theories produce predictions that are fundamentally different from the output from the head nods detector. On the one hand, most of the detected head nods are not accounted by any of the theories. On the other hand, most of the predictions by the theories cannot be paired to a detected head nod. There are a few ways to interpret this. Either the theories do not capture head movement and nods sufficiently, or, the detected head nods do not align with our intuitive notion of a head nod. Realistically, the answer is somewhat a mixture of both.

Chapter 6 originates from the attempt to validate the automatic head nods detector. The decision to explore this avenue, as opposed to searching for theories that capture a larger percentage of the detected head nods, stems from the desire to make sure head nods and their characteristics are well defined in this work. Otherwise, the meaning of the comparison between the theories can be questionable. For example, imagine that the head nods detector would

have been found to systematically trigger only when speakers use their head to emphasise their speech. This would affect the interpretation of the comparison between theories significantly. So, from that point forward, the focus of this work is shifted from the comparison between existing theories to a somewhat lower level interest in the motion and semantic characteristics of head nods, as perceived by annotators, in comparison to motion capture based detector. The naive head nods detector employed throughout this work reflects the idea that a head nod is a movement downwards then upwards of the head. It seems that this overly simplified assumption is common in the literature. When tested a more nuanced picture is revealed. First, the agreement on what is a head nod is moderate, even when annotators are presented with video snippets that include only automatically detected nods. Second, head nods share movement characteristics with other behaviours, specifically laughter. Lastly, head nods are more accurately defined by their semantic characteristics than by their movement properties, suggesting that future detectors should incorporate more contextual features than movement alone. In addition, and although unintentionally, this chapter provides partial answers to the questions raised in the previous chapter regarding the adequacy of existing theories to explain head movement in conversation. Relying on standard annotation methods, the clustering analysis generally fails to break down speakers' behaviours into discrete recognisable groups. This implies that our vocabulary to describe speakers' non-verbal behaviour is generally lacking.

Chapter 6 concludes that the agreement on what is considered a head nod is moderate at best and that their movement characteristics are not as well defined as naively thought. Does this mean that head nods are not a useful construct for the analysis and understanding of head movement? In my view the answer to this question is no. Head nods found to have clear semantic meaning. They are associated with listeners, give feedback, and show acceptance. Future work that aim to automatically detect head nods should therefore develop ways to understand the conversational context better, rather than invest in movement processing alone. This might include exploring co-occurrences of head nods and verbal responses, linguistic features, and deeper understanding of turn taking structures.

When looking at this work as a whole an interesting pattern emerge. Whereas it seems natural for a scientific work to suggest further research that relies on its findings and builds on top of them, this thesis generally follows a different theme. The relationship between each two subsequent chapters is mainly of unravelling assumptions. Each chapter finishes by highlighting problems with its implied expectations, and is followed by a chapter that tries to simplify the problem to answer a more basic question. This, in a sense, tells something about our

understanding of head movement, head nods, and the literature around these. More often than not, the descriptions of head nods in the literature are rather simplistic. Head movement, in general, is partially explained by multiple theories, that are rarely compared. This opens a lot of room for future research, while at the same time does not provide solid enough ground to rely on.

This observation does not aim to disprove the evidence around existing theories and their predictions in regards to head movement. For example, this work does not claim to disprove the existence of head movement mimicry, nor supply additional evidence to support it. What it tries to say is that this theory, and the other ones explored here, are perhaps minor effects in the general scheme of how non-verbal cues are used in free conversation. Hopefully, future research will delve deeper into the nature of head nods, find better ways to define them and contextualise them. This can help the development of future theories that might, eventually, be able to explain a larger proportion of the varied behaviours we observe in free conversation.

Bibliography

Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, pages 7–22. Copenhagen.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., and Turk, M. (2004). Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 13(4):428–441.

Bailenson, J. N. and Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819.

Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941–952.

Bavelas, J. B., Coates, L., and Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580.

Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., and Petit, J.-P. (1997). Itu-t recommendation g. 729 annex b: a silence compression

scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9):64–73.

Bertrand, A. and Moonen, M. (2010). Energy-based multi-speaker voice activity detection with an ad hoc microphone array. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 85–88. IEEE.

Blache, P., Bertrand, R., and Ferré, G. (2009). Creating and exploiting multimodal annotated corpora: The ToMA project. In *Multimodal Corpora*, pages 38–53. Springer Berlin Heidelberg.

Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Boholm, M. and Allwood, J. (2010). Repeated head movements, their function and relation to speech. In *Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analysing multimodality*, pages 6–10. Citeseer.

Boletsis, C. and Cedergren, J. E. (2019). Vr locomotion in the new era of virtual reality: an empirical comparison of prevalent techniques. *Advances in Human-Computer Interaction*, 2019.

Brady, H. E. (2011). *Causation and Explanation in Social Science*. Oxford University Press.

Cerrato, L. and Svanfeldt, G. (2006). A method for the detection of communicative head nods in expressive speech. In *The Second Nordic Conference on Multimodal Communication, Göteborg, 07/04/2005*, pages 153–165. Göteborg University.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893–910.

Chartrand, T. L. and Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.

Chen, L. and Harper, M. P. (2009). Multimodal floor control shift detection. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 15–22. ACM.

Chen, Y., Yu, Y., and Odobez, J.-M. (2015). Head nod detection from a full 3d model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 136–144.

De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

de Kok, I. (2013). *Listening Heads*. PhD thesis, University of Twente. SIKS Dissertation Series No. 2013-29.

de Kok, I. and Heylen, D. (2012). Integrating backchannel prediction models into embodied conversational agents. In *Intelligent Virtual Agents*, pages 268–274. Springer.

Dong, L., Jin, Y., Tao, L., and Xu, G. (2007). Recognition of multi-pose head gestures in human conversations. In *Fourth International Conference on Image and Graphics (ICIG 2007)*. IEEE.

Duncan, S., Brunner, L. J., and Fiske, D. W. (1979). Strategy signals in face-to-face interaction. *Journal of Personality and Social Psychology*, 37(2):301.

Edelsky, C. (1981). Who's got the floor? *Language in society*, 10(3):383–421.

Ferrer, L., Shriberg, E., and Stolcke, A. (2002). Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *Seventh International Conference on Spoken Language Processing*.

Fujie, S., Matsuyama, Y., Taniyama, H., and Kobayashi, T. (2009). Conversation robot participating in and activating a group communication. In *Tenth Annual Conference of the International Speech Communication Association*.

Gepp, M. (2017). Roomscale 101 – an introduction to roomscale vr. `https://blog.vive.com/us/2017/10/25/roomscale-101/`. Accessed 2021-08-12.

Giannoulis, D., Massberg, M., and Reiss, J. D. (2013). Parameter automation in a dynamic range compressor. *Journal of the Audio Engineering Society*, 61(10):716–726.

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, pages 97–121.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., and Morency, L.-P. (2006). Virtual rapport. In *IVA*, volume 6, pages 14–27. Springer.

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer.

Griffin, H. J., Aung, M. S., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., and Bianchi-Berthouze, N. (2013). Laughter type recognition from whole body motion. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 349–355. IEEE.

Gurion, T., Healey, P. G., and Hough, J. (2018). Real-time testing of non-verbal interaction: An experimental method and platform. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Aix-en-Provence, France. SEMDIAL.

Gurion, T., Healey, P. G., and Hough, J. (2020). Comparing models of speakers' and listeners' head nods. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*, Whaltham, MA. SEMDIAL.

Hadar, U., Steiner, T. J., Grant, E. C., and Clifford Rose, F. (1983a). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.

Hadar, U., Steiner, T. J., Grant, E. C., and Clifford Rose, F. (1983b). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46.

Hale, J. and Hamilton, A. F. d. C. (2016). Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific reports*, 6.

Hale, J., Ward, J. A., Buccheri, F., Oliver, D., and de C. Hamilton, A. F. (2019). Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of Nonverbal Behavior*, 44(1):63–83.

Hasler, B. S., Hirschberger, G., Shani-Sherman, T., and Friedman, D. A. (2014). Virtual peacemakers: Mimicry increases empathy in simulated contact with virtual outgroup members. *Cyberpsychology, Behavior, and Social Networking*, 17(12):766–771.

Hayamizu, S., Hasegawa, O., Itou, K., Sakaue, K., Tanaka, K., Nagaya, S., Nakazawa, M., Endoh, T., Togawa, F., Sakamoto, K., et al. (1996). Rwc multimodal database for interactions by integration of spoken language and visual information. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2171–2174. IEEE.

Healey, P. G., Lavelle, M., Howes, C., Battersby, S. A., and McCabe, R. (2013). How listeners respond to speaker's troubles. In *CogSci*, pages 2506–2511.

Healey, P. G., Plant, N., Howes, C., and Lavelle, M. (2015). When words fail: collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series,(Chicago)*.

Hepburn, A. and Bolden, G. B. (2013). The conversation analytic approach to transcription. *The handbook of conversation analysis*, pages 57–76.

Holler, J., Kendrick, K. H., Casillas, M., and Levinson, S. C. (2016). *Turn-taking in human communicative interaction*. Frontiers Media.

Hough, J. and Schlangen, D. (2017). Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., and Ginzburg, J. (2016). Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

Howes, C., Healey, P. G., Purver, M., and Eshghi, A. (2012). Finishing each other's... responding to incomplete contributions in dialogue. In *Proceedings of the Cognitive Science Society*, volume 34.

Huang, L., Morency, L.-P., and Gratch, J. (2011). Virtual rapport 2.0. In *Intelligent Virtual Agents*, pages 68–79. Springer.

Huang, Y., Vinyals, O., Friedland, G., Muller, C., Mirghafori, N., and Wooters, C. (2007). A fast-match approach for robust, faster than real-time speaker diarization. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 693–698. IEEE.

Ivanov, A. V. and Riccardi, G. (2010). Automatic turn segmentation in spoken conversations. In *Eleventh Annual Conference of the International Speech Communication Association*.

Kang, Y. G., Joo, H. J., and Rhee, P. K. (2006). Real time head nod and shake detection using HMMs. In *Lecture Notes in Computer Science*, pages 707–714. Springer Berlin Heidelberg.

Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM.

Karpiński, M., Jarmołowicz-Nowikow, E., and Czoska, A. (2015). Gesture annotation scheme development and application for entrainment analysis in task-oriented dialogues in diverse cultures. In *Proceedings of GESPIN 2015 Conference, Nantes, France*, pages 161–166.

Kawato, S. and Ohya, J. (2000). Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE Comput. Soc.

Kousidis, S., Malisz, Z., Wagner, P., and Schlangen, D. (2013). Exploring annotation of head gesture forms in spontaneous human interaction. In *Proceedings of the Tilburg Gesture Meeting (TiGeR 2013)*.

Kuhlen, A. K. and Brennan, S. E. (2012). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, 20(1):54–72.

Lakin, J. L. and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science*, 14(4):334–339.

Lala, D., Inoue, K., and Kawahara, T. (2018). Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*. ACM Press.

Lala, D., Inoue, K., and Kawahara, T. (2019). Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*, pages 226–234. ACM.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Leander, N. P., Chartrand, T. L., and Bargh, J. A. (2012). You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological science*, 23(7):772–779.

Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer.

Maatman, R., Gratch, J., and Marsella, S. (2005). Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*, pages 25–36. Springer.

Maier, A., Hough, J., and Schlangen, D. (2017). Towards deep end-of-turn prediction for situated spoken dialogue systems. *Proceedings of INTERSPEECH 2017*.

McCabe, R., Healey, P. G., Priebe, S., Lavelle, M., Dodwell, D., Laugharne, R., Snell, A., and Bremner, S. (2013). Shared understanding in psychiatrist–patient communication: Association with treatment adherence in schizophrenia. *Patient Education and Counseling*, 93(1):73–79.

Meteer, M. W. (1995). *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania.

Morency, L.-P., de Kok, I., and Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. Springer.

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces - ICMI '05*. ACM Press.

Navarretta, C. (2011). Annotating non-verbal behaviours in informal interactions. In Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., and Nijholt, A., editors, *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pages 309–315, Berlin, Heidelberg. Springer Berlin Heidelberg.

Nguyen, L., Odobez, J.-M., and Gatica-Perez, D. (2012). Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 289–292. ACM.

Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.

Nishimura, R., Kitaoka, N., and Nakagawa, S. (2007). A spoken dialog system for chat-like conversations considering response timing. In *Text, Speech and Dialogue*, pages 599–606. Springer.

Paggio, P. and Navarretta, C. (2010). Feedback in head gestures and speech. In *LREC 2010 Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pages 1–4. Citeseer.

Paggio, P. and Navarretta, C. (2012). Classifying the feedback function of head movements and face expressions. In *Multimodal Corpora: How Should*

Multimodal Corpora Deal with the Situation? Workshop Programme, pages 34–37.

Poggi, I., D'Errico, F., and Vincze, L. (2010). Types of nods. the polysemy of a social signal. In *LREC*.

Poppe, R., Truong, K., Reidsma, D., and Heylen, D. (2010). Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents*, pages 146–158. Springer.

Puckette, M. S. (1996). Pure data: another integrated computer music environment. *Proceedings of the second intercollege computer music concerts*, pages 37–41.

Roddy, M., Skantze, G., and Harte, N. (2018). Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461*.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97(5):1295–1345.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Schlangen, D. (2006). From reaction to prediction: Experiments with computational models of turn-taking. In *Ninth International Conference on Spoken Language Processing*.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, Berkeley.

Shriberg, E. E., Bates, R. A., and Stolcke, A. (1997). A prosody only decision-tree model for disfluency detection. In *Eurospeech*, volume 97, pages 2383–2386.

Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.

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

Tiedens, L. Z. and Fragale, A. R. (2003). Power moves: complementarity in dominant and submissive nonverbal behavior. *Journal of personality and social psychology*, 84(3):558.

Verberne, F. M., Ham, J., Ponnada, A., and Midden, C. J. (2013). Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In *International Conference on Persuasive Technology*, pages 234–245. Springer.

Vrijsen, J. N., Lange, W.-G., Becker, E. S., and Rinck, M. (2010a). Socially anxious individuals lack unintentional mimicry. *Behaviour Research and Therapy*, 48(6):561–564.

Vrijsen, J. N., Lange, W.-G., Dotsch, R., Wigboldus, D. H., and Rinck, M. (2010b). How do socially anxious women evaluate mimicry? a virtual reality study. *Cognition and Emotion*, 24(5):840–847.

Wales, D. J. and Doye, J. P. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116.

Wallis, T. (2016). What is social vr? `https://www.vr-intelligence.com/social-vr-101`. Accessed 2018-09-04.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Wu, K., Liu, C., and Calvo, R. A. (2020). Automatic nonverbal mimicry detection and analysis in medical video consultations. *International Journal of Human-Computer Interaction*, 36(14):1379–1392.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Appendix A

Head nod detection pseudo code

```
# Thresholds in millimetres
nodding = -4
not_nodding = 2

# Are we ready to detect nods?
ready = true
# Interpolator to up-sample from arbitrary frequency to 100Hz
interpolator = Interpolator(freq=100)
# 2nd order Butterworth filters
lowpass = LowPassFilter(freq=4)
highpass = HighPassFilter(freq=1)

# This function is called by the application at varying intervals
def process(head_vertical_position):
  now = time.now()
  for x in interpolator.interpolate(now, head_vertical_position):
    x = lowpass.filter(x);
    x = highpass.filter(x);
    if ready and x < nodding:
      ready = false;
      # HEAD NOD DETECTED!
    elif -nod_nodding < x < not_nodding:
      ready = true;
```

Appendix B

Apartment design task

Imagine the following: You've been given substantial resources (let's say, £500k) to use on an apartment that you will share with your conversational partner. There are only the constraints that there can only be one kitchen (which will be shared), one bathroom (which will be shared), and one living room (which will be shared). Your task now is to plan this apartment together, both the shared and private areas as well as furniture, decoration, things like TV etc. You have 15 minutes for this task.

Appendix C

Questionnaires

Note that some questionnaires include open questions. Space to write an elaborated answer is removed in this appendix for brevity. The questionnaires given to the participants include space for a paragraph per open question.

C.1 Demographic questionnaire

- What is your age? _____
- What gender do you identify with?
  - Male
  - Female
  - Other _____
  - Prefer not to say

C.2 Familiarity with virtual reality questionnaire

- Have you experienced virtual reality with a head mounted display?
  - Never
  - Once
  - A few times
  - Regularly
- Have you experienced virtual reality with a room scale feature (the ability to freely walk in the virtual reality environment as opposed to just standing or sitting)?

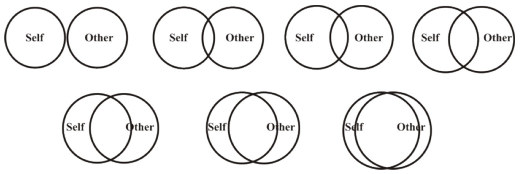
  - Never
  - Once
  - A few times
  - Regularly

- Have you experienced social virtual reality (hanging out with other people in a shared virtual environment)?
  - Never
  - Once
  - A few times
  - Regularly

C.3 Social closeness questionnaire

Please answer the following questions considering your experimental partner

- How many years have you known your experimental partner? _____
- How often are you in contact with your experimental partner?
  - Never
  - Occasionally
  - A few times a month
  - A few times a week
  - Daily
- Please circle the picture which best describes your relationship.



C.4 Post-experience questionnaire

- What were your strategies to maximize your score?
- What were your strategies to detect that someone is faking attention?
- Any other comment?

Appendix D

Head and torso movement annotation scheme

This manual explains how to annotate movement in short video snippets containing dialogues between two participants. The annotations should be filled in the supplied spreadsheet, with each video snippet corresponding to one row in the spreadsheet.

The filename of the video snippet to annotate is indicated in the `Filename` column in the spreadsheet. Only one participant should be annotated per row, as indicated in the `Participant` column in the spreadsheet. If the value is `Left`, fill in the relevant information about the participant on the left side of the video frame. If the value is `Right`, you guessed it...

The video snippets are 6 seconds long, and might contain several movements. Your goal is to annotate the one closest to the middle time point of the snippet. First, watch the snippet a few times to decide what movement to annotate. Fill in the columns, left to right, following the instructions below. Unless otherwise noted a column can be left empty, and the values in a column are not expected to be evenly balanced (e.g. an hypothetical column `C` may have a lot more video snippets annotated with the value `x` than with `y`).

D.1 `Feedback`

D.1.1 `Direction`

This column indicates whether the motion is meant to `give` feedback, or to `elicit` feedback. Fill in one of these values if that is the participant's aim, in your opinion, in producing this motion.

D.1.2 `Acceptance`

This column indicates whether the motion signals acceptance (fill in `accept`) or rejection (fill in `reject`). Fill in one of these values if that is the participant's aim, in your opinion, in producing this motion.

D.2 `Movement`

If you chose to fill in any of the `Movement` sub-columns make sure to fill in all of them.

D.2.1 `Body part`

This column indicates whether the movement is mainly a `head` movement or a `torso` movement. Other body parts should be ignored. Your decision should be based on the body part that describes the movement better. If in doubt bias towards `torso` movement. Leave this column empty if what you see in the video snippet cannot be described as a `head` or `torso` movement.

D.2.2 `Sagittal`

Coming from anatomy, the sagittal plane is perpendicular to the ground, separating the body left from right. This column should be filled in with either `yes` or `no`. A value of `yes` indicates that the movement is contained within the sagittal plane, while a value of `no` indicates that the movement is outside of the sagittal plane. Some examples of head movements are depicted in Figure D.1. The top row shows movements that should be annotated with `yes` in the `Sagittal` column. The bottom row shows movements that should be annotated with `no`. Note that sagittal and non-sagittal movement can describe displacement (change in head or torso position) as well as rotation.

D.2.3 `Repetitions`

This column indicates, qualitatively, how many repetitions happen in the observed movement. The possible values for this column are:

`partial` means that the movement does not complete a full circle. For example: moving the head from one side to the other, once, without returning it to the original position.

`single` means that the movement returns to its original position or rotation. For example: a single head nod.

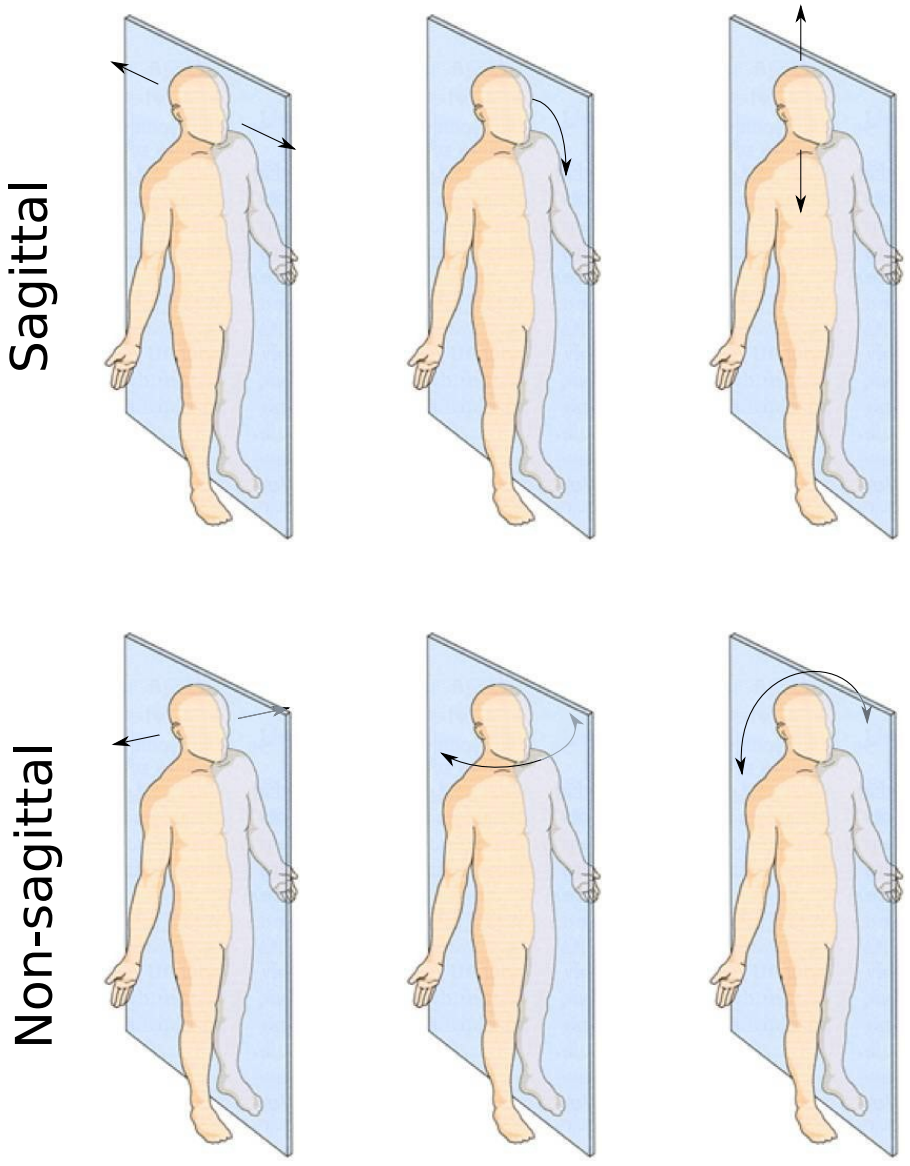


Figure D.1: Examples of sagittal (top row) and non-saggital (bottom row) movements.

`multiple` means that the movement repeats more than once. For example: multiple shakes of the head from left to right as done while displaying disagreement.

D.3 Semiotic function

This column indicates the communicative nature of the movement. When filling in this column think about the motion you choose to annotate. Ignore hand gestures and utterances that do not coincide that movement (e.g. a thumbs up gesture immediately after the head motion you are annotating). Note that head nods and *backchannels* that indicate, for example, the listener's attention / acceptance, do **not** fall into any of these categories. This manual is supplied with one video snippet exemplifying each semiotic function. The possible values for this column are:

`deictic` for indexical deictic gestures that locate aspects of the discourse in the physical space, for example by pointing. They can be used to address people. They can refer to objects within the room, or to hypothetical objects that could have been in the room.

`non-deictic` for indexical non-deictic gestures that indicate a causal relation between the gesture and the effect it establishes. The small movements that accompany speech and underline its rhythm, and that some people have called *batonic* or *beat* gestures, fall into this category.

`iconic` gestures (including so-called metaphoric gestures) express some semantic feature by similarity or homomorphism. Examples are gestures done with two hands to comment on the size (length, height, etc.) of an object mentioned in the discourse.

`symbolic` gestures (emblems) are gestures in which the relation between form and content is based on social convention (e.g. the okay gesture). They are culture-specific.

D.4 Binary features

The last few columns should be filled with either `yes` or `no`. **Do not** leave any of these columns empty.

`Head nod` : whether or not the movement is, or contains, a head nod.

`Laughter` : whether or not the participant is laughing.

`Face touching` : whether or not the participant touches their face or head tracking sensor (attached to the cap).

`Distraction` : whether the motion you observe is a result of something external to the conversation. For example, when the participant turns to approach the experimenter, receives a phone call and pulls their phone from their pocket, rotates to check some noises from outside the room, etc.

`Technical issue` : whether you think there is no movement in the video snippet and it is provided for annotation by mistake.