Comparing models of speakers' and listeners' head nods

Tom Gurion, Patrick G.T. Healey and Julian Hough
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{t.gurion,p.healey,j.hough}@qmul.ac.uk

Abstract

Head movement, and head nods in particular, are important communicative signals in face to face conversations. Listeners' head nods can be characterised as acknowledgements to the speaker or as showing support when the speaker encounters problems in completing their turn. Speakers' head movement, on the other hand, is often explained as mimicry. These different explanations of the same observed phenomena are rarely compared, and often contradict each other. This study compares models of head nods in free conversation based on three theories: mimicry, backchannels, and listener responses to speaker trouble. Pairs of participants freely conversed for 15 minutes while fitted with a motion capture system. Head nods where automatically extracted from the data, and compared to predictions by the models. The results highlight the differences between head nod patterns of speakers and listeners, and suggest that the theories are not mutually exclusive. Comparing the theories yield inconclusive results and fail to determine which one best explains head nods. Lastly, the vast majority of the observed head nods are not accounted by any of the investigated models, implying that mimicry, backchannels, and responses to speech disfluencies alone cannot explain head nods exhaustively.

Face-to-face conversations are multimodal in nature. Verbal and non-verbal communication both contribute to the content of the conversation and coordinate its structure. Research of non-verbal communication shows a number of coordinating phenomena: listeners gesture to demonstrate their attendance to the speaker (Goffman, 1955); speakers respond to listeners gaze, and often change who they are looking at depending on eye-contact with their addressees (Goodwin, 1979); when listeners fail to provide the expected listening behaviours speakers get confused (Bavelas et al., 2000).

This study aims to compare three theories that explain head nods in conversation: nods as backchannels, nods as mimicry, and nods as listener responses to speaker trouble (specifically, disfluency). There are many studies about backchannels and mimicry, but to our knowledge, comparisons of these as competing explanations for the same observed phenomena are rare. Whereas two of these theories (backchannels and responses to speaker trouble) are specific to listeners, mimicry theory is agnostic to conversational roles. In other words, the mimicry literature in general ignores speaker-hearer roles. Whereas some studies investigate how speakers mimic listeners (Bailenson and Yee, 2005; Vrijsen et al., 2010) others look at listener mimicry of speakers (Maatman et al., 2005; Leander et al., 2012; Hale et al., 2018), while both apply the same reasoning and theory.

In this study we aim to test how well these theories explain observed head nods in free conversations in respect to speaker-hearer roles. We also address the question of overlap between the theories, and whether or not they provide an adequate account for head nods.

We conducted a study with pairs of participants having free dialogues, while fitted with a motion caption system. Their head movement was analysed for head nods and these were compared to the predicted behaviour according to models based on the three theories above.

The rest of the paper is as follows: Section 1 overviews the theoretical and empirical basis of the models of head nods in dialogue we implement; Section 2 describes our interactive head nod prediction models; Section 3 describes the methods we use for data collection and the automatic techniques we use for testing our model; Section 4 gives and discusses the results of applying our models to our data, before we conclude in Section 5.

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020, Online, hosted from Massachusetts, USA.

1 Related work

1.1 Backchannels

Natural and engaging conversations rely on the ability of the interlocutors to achieve common ground. This process often involves feedback from listeners for their understanding, or misunderstanding. Timed listener responses, known as backchannels (Yngve, 1970), can be non-verbal (e.g. head nods), or para lingual (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).

Surface features like speaker-listener eye contact and speech prosody are often enough to model backchannels. Ward and Tsukahara (2000) suggeste a backchannel prediction model that is based on the speaker's prosody alone. Their model, summarised in Table 1, provides a set of simple hand-crafted rules that use the speaker "vocalisation state" (speaking or silent) and pitch, to predict listeners' backchannels. Later studies often rely 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, backchannel predictor (Morency et al., 2008; Poppe et al., 2010).

1.2 Mimicry

A well known hypothesis is that conversational partners often mimic each other's postures, gestures, use of language, facial expression, and more. It is commonly claimed that mimicry in conversation has persuasive power and can increase likeability, empathy, and feeling of closeness (for a review see Chartrand and Lakin, 2013).

The automatic tendency to imitate others in social interactions is often described as the "chameleon effect" (Chartrand and Bargh, 1999). It is usually explained by the perception-behaviour link: a cognitive mechanism that link together perception and action and suggests that we have to act in order to perceive. This implies that mimicry is an unconscious automatic behaviour that is a by product of our ability to process social encounters.

The temporal properties of mimicry have been studied extensively. Early research in mimicry with virtual reality reported that head movement mirroring by a virtual agent, with a 4 second lag, increases rapport (Bailenson and Yee, 2005). A recent study (Hale et al., 2018) challenged this 4 second mimicry lag from Bailenson's work (Bailenson et al., 2004; Bailenson and Yee, 2005). The authors highlight that a 4 second lag implies that mimicry is not a reactive process and it relies on memory to operate, which is rather unlikely. One alternative model describes mimicry as a predictive process, similar to the mechanism that allows musicians to coordinate their playing, and implies no time lag. The third alternative 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 to measure the head pitch in dyadic conversations and found that listeners' low-frequency head movement follows the speaker's head movement after a 600 milliseconds delay.

1.3 Responses to speaker's troubles and disfluencies

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, as happen commonly in free conversations (Schegloff, 1992).

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.

A study exploring head movement after disfluent utterances found that addressees tend to nod more between 1 to 3 second after a disfluency (Healey et al., 2013). In addition, listeners gesture more when the speaker is disfluent, and their gestures tend to become similar to those of the speaker (Healey et al., 2015).

2 Models implementation

2.1 Backchannels

Ward and Tsukahara (2000) suggest a prosody-based backchannels prediction model. This model, summarised in Table 1, provides a set of simple hand-crafted rules that use the pitch of the speaker's voice, and the timing they started or stopped speaking, to predict hearers' backchannels. It is important to highlight that the original model was designed to predict verbal backchannels. Neverthe-

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 1: Ward and Tsukahara (2000) backchannels prediction algorithm (copied from Poppe et al., 2010).

less, several studies have used it to predict non-verbal backchannels, and specifically head nods (e.g. Morency et al., 2008; Poppe et al., 2010), based on the assumption that verbal and non-verbal backchannels are generally interchangable (Bavelas et al., 2000).

This model was chosen for its simplicity and ease of interpretation. While machine learning based models often out-perform it (e.g. Morency et al., 2008), they are usually harder to interpret. The interpretability 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 1). 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 was 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)[1] and utilises the voice activation detection application py-webrtcvad[2] to decide if there is speech in the audio data or not (rule P3 in table 1). One instance of the model is run for each participant. It analyses the audio stream from the participant's microphone and predicts backchannels for their conversational partner.

[1]The backchannels detection server is available online at https://github.com/Nagasaki45/backchannels.

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

2.2 Mimicry

To tackle the question of mimicry lag two mimicry models were implemented. They predict head nods for one participant 600 milliseconds (following Hale et al., 2018) and 4 seconds (following Bailenson and Yee, 2005) after a head nod by their partner. These models rely on a head nod detector that is described in detail in Section 3.2.

Note that the mimicry models use head nod mimicry, and not head movement mimicry as most mimicry studies suggested. There are two reasons for this decision: firstly, we aim to compare mimicry and backchannels as possible explanations for communicative head movement. Therefore, to obtain a meaningful comparison, both models should operate on the same physical movement. Second, the mimicry 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 mimicry research suggests (Chartrand and Bargh, 1999), 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, Hale et al. (2018) concluded that "the cognitive mechanisms generating mimicry of head nods act with a constant lag of around 0.588 msec".

2.3 Disfluency

Speech disfluencies are detected using IBM Watson's speech-to-text service, and the "deep_disfluency" software, a state-of-the-art incremental model that tags disfluencies in transcripts (Hough and Schlangen, 2017). The audio from each participant's microphone is sent to IBM Watson. The returned transcription is passed through the deep disfluency tagger. Interregnum and edit term tags (<e/>) and repair start tags (<rpS/>) are interpreted as disfluencies.

3 Methods

3.1 Participants

Thirteen pairs of native English speakers that knew each other in advance (14 women, 12 men), age 18-26 (mean: 20.8, std: 1.9) participated in the study. 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.

3.2 Apparatus

Two participants at a time participated in the 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.

To track their head movement the participants wear baseball caps fitted with HTC Vive trackers[3] on their visors. They also held HTC Vive handheld controllers[4] 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. The logged movement data includes the head and hands position (Cartesian X, Y, and Z values), and rotation (Euler angles X, Y, and Z), sampled at the frame rate of the motion data capture application[5], which varies between 60Hz and 90Hz.

Head nods are detected based on the vertical position of the head tracker. To guarantee a constant rate, the vertical head position is up-sampled, linearly, to 100Hz. 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 above study processed data at 60 frames per second, that is 0.18 millimetre per millisecond, or, translated to our 100hz sample rate, 1.8 millimetres per sample. This value, of 1.8 millimetres per sample, seemed too sensitive. We fine tuned it manually until the system was sensitive enough to detect most head nods but not too sensitive to be triggered by other movement.

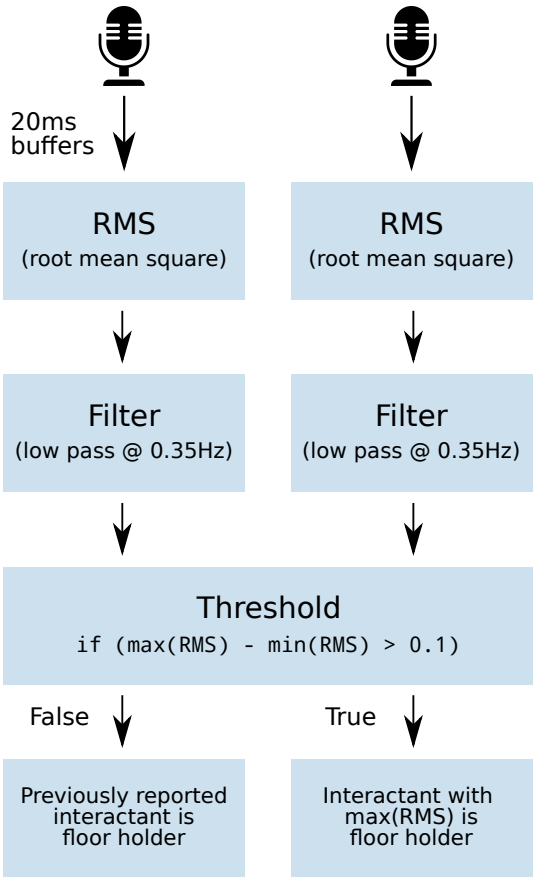

Figure 1: Schematic diagram of the floor control detection model.

Recording the participants' voices in one room 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 iZotope RX 6 De-bleed[6] software was used 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 was not tested exhaustively, but a few measurements suggest that the bleed was reduced by up to 40dB. This version of the audio files was used instead of the original audio files for the floor control detection model (described next), and the backchannels and disfluency models.

A simple and accurate floor control detection

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

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

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

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

model (Gurion et al., Unpublished) processed the audio from the participants' microphones to determine the floor holder at any given moment. This model, presented in Figure 1, relies on simple audio processing techniques of filters and thresholds to operate. Audio from the microphones of the two interactants is processed in buffers of 20 milliseconds. For each buffer the root mean square (RMS) value is calculated. These values are filtered by low pass filters with a cutoff frequency of $0.35Hz$. If the difference between the minimal and the maximal filtered RMS values is larger than 0.1 the interactant with the maximal filtered RMS value is the floor-holder. Otherwise the previously reported floor-holder is the floor-holder again.

Lastly, the backchannels, mimicry, and disfluency models described earlier logged predictions for partner nods.

3.3 Procedure

First, the participants filled a demographic questionnaire, followed by a questionnaire about their social relationship. 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 participants were then fitted with the motion capture system, discussed the task for 15 minutes, until the experimenter asked them to stop.

4 Results and discussion

4.1 Head nods frequency

First, head nods frequency was calculated separately for floor holders and non floor holders. For each participant, the number of head nods while holding the floor was divided by the total time they held the floor to find the nodding frequency. Nodding frequency for non floor holders was calculated in a similar fashion. Comparing these values suggest that floor holders nod more frequently (M=0.27Hz, SD=0.14) than non floor holders (M: 0.18Hz, SD=0.09; $t(25) = 4.11$, $p < 0.001$).

4.2 Overlap between models

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

To test that we calculated the overlap for every combination of 2, 3, or all 4 models. 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. an XNOR logical 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 2 shows the difference between the overlap and the overlap expected by chance for all combinations of models. Each dot indicate 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 $\alpha < \frac{0.05}{11}$).

The significant overlap between the backchannels and the disfluency models can be interpreted in a few ways. First, these models 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 shares 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 backchannels.

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).

We couldn't find a theoretical explanation, nor an explanation based on the models implementation, for the higher than chance overlap for the mimicry600 $\cap$ mimicry4000 combination. Lastly, given the overlaps found between combinations of

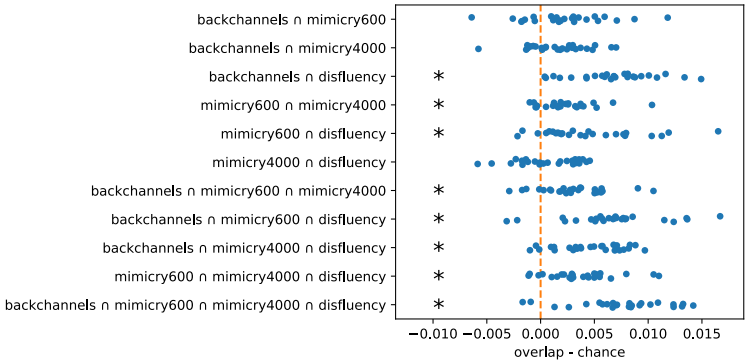

Figure 2: The difference between overlap and overlap expected by chance for combinations of models, per participant. Model combinations with overlap significantly higher than chance are marked with an asterisk.

two models, it is not surprising that combinations of 3 and 4 models also overlap.

4.3 Models comparison

We used precision, recall, and F1-scores (the harmonic mean of precision and recall), per participant, to assess how well the models perform. Because the models' predictions and the detected head nods are timed events we had to adapt the usual precision and recall definitions 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 head nod was 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) we believe that this method is appropriate.

Figure 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 4.35 seconds. 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 4 can help explaining the differences in recall values between the models. The left bar graph

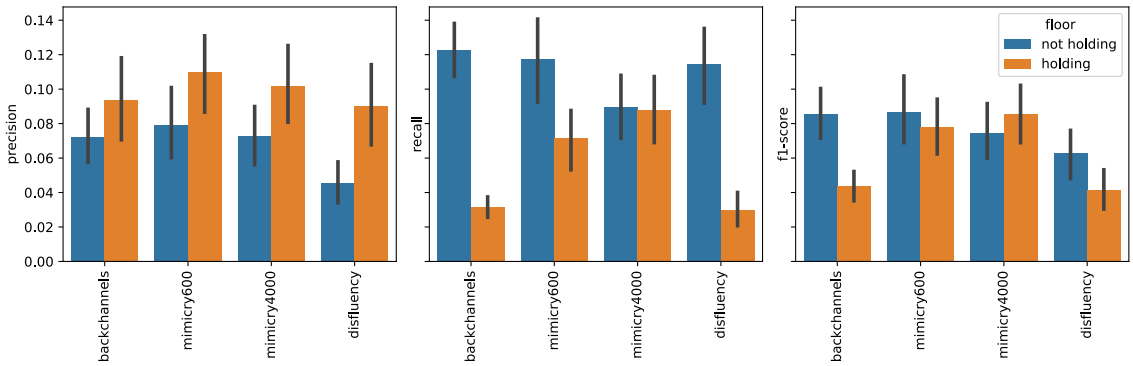

Figure 3: Precision, recall, and F1-score for the models' predictions for participants holding or not holding the floor. The error bards indicate a 95% confidence interval of the mean.

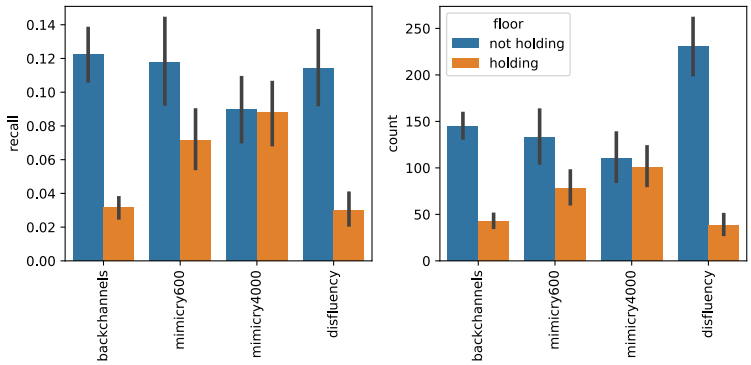

Figure 4: Recall performance of each model (left) and number of predictions generated by each model per participant, when holding and not holding the floor (right). The error bards indicate a 95% confidence interval of the mean.

shows the recall values of the models, for floor holders and non floor holders (same as the middle panel of Figure 3), while the right bar graph show 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.

4.4 Accounted head nods

Here we assess 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 recall across all participants is 0.29. This low value indicates that the vast majority of the detected head nods are not accounted by any of the models we investigated. The recall for floor holders is 0.20 and for non floor holders is 0.38. As two out of the four models, the backchannels and disfluency models, are desgined specifically for listeners, the union recall for non floor holders is expected to be higher. Nevertheless, even for non floor holders most of the 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 backchannels research (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 section. Lastly, although the models discussed here cover a few different theories that are extensively studied, we are sure 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 nodding behaviour. A theory that specifically deals with speakers' head movement and nods might shed a light on the head nods that our models failed to predict.

4.5 Limitations

This study relies on an automatic head nod detector. This detector was never validated properly and it seems that it is a source of significant noise and inaccuracies in the analysis. Even on the theoretical level, there is no agreement on what head movements should be considered a head nod (e.g. is it related to pitch or height of head). Further research is required to either come up with a verified model or, perhaps, a broader definition of head nods that fits a concrete form of head movement.

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. The method looks for predictions around detected nods to estimate true positive, and derive the precision and recall from it. It was used before for similar task (Poppe et al., 2010), to our knowledge the method, and the window choice, were never properly evaluated.

5 Conclusion

This study compares predictions of models inspired by three theories to head nods of floor holders and non floor holders observed in free dialogues as annotated automatically by a head nod detector. According to our detector, the head nod frequency of floor holders is higher than that of non floor holders. We discover 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. Whereas some theories achieve higher recall for specific conversational roles the difference is usually a result of more predictions generated for these roles. Lastly, most of the observed head nods in the data went undetected by any of the models we investigated. This can be a result of technically insufficient methods of detecting head nods and matching these with predictions. Alternatively, it might suggest that the theories we investigate here are not enough to explain head nods by speakers and listeners in free conversation.

Acknowledgment

This work was supported by grant EP/L01632X/1 for the Centre for Doctoral Training in Media and Art Technology (MAT http://www.mat.qmul.ac.uk/) from the UK Engineering and Physical Sciences Research Council (EPSRC) and the UK Arts and Humanities Research Council (AHRC).

References

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

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

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941–952.

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

Tanya L Chartrand and Jessica L Lakin. 2013. The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, 64:285–308.

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

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

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

Tom Gurion, Julian Hough, Sophie Skach, and Patrick G.T. Healey. Unpublished. A simple, real-time model of turn-taking in dialogue. Manuscript in preparation.

Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. 2018. Are you on my wavelength? interpersonal coordination in naturalistic conversations. Preprint available at https://psyarxiv.com/5r4mj/.

Patrick GT Healey, Mary Lavelle, Christine Howes, Stuart Adam Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles. In *CogSci*, pages 2506–2511.

Patrick GT Healey, Nicola Plant, Christine Howes, and Mary Lavelle. 2015. When words fail: collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series,(Chicago)*.

Julian Hough and David Schlangen. 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)*.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

Iwan de Kok. 2013. *Listening Heads*. Ph.D. thesis, University of Twente. SIKS Dissertation Series No. 2013-29.

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

RM Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*, pages 25–36. Springer.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. Springer.

Ronald Poppe, Khiet Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents*, pages 146–158. Springer.

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

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

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Berkeley.

Janna N Vrijsen, Wolf-Gero Lange, Ron Dotsch, Daniël HJ Wigboldus, and Mike Rinck. 2010. How do socially anxious women evaluate mimicry? a virtual reality study. *Cognition and Emotion*, 24(5):840–847.

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

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