**About backchannels, rapport and turn-taking**

(including refs to literature and summary of previous virtual rapport agents by ICT)

**Backchanneling theory**

* <https://en.wikipedia.org/wiki/Backchannel_(linguistics)>
* backchannel response can be [verbal](https://en.wikipedia.org/wiki/Language), [non-verbal](https://en.wikipedia.org/wiki/Non-verbal), or both
* signifying the listener's attention, understanding, or agreement, rather than conveying significant information
* Verbal:
  + "yeah", "uh-huh", "hmm", and "right"
* For example, backchannel responses are not only a key component of spoken languages but they are also important in sign languages.[7]
* Another example of the variability that occurs in backchannel responses in different languages is that Germans produce less backchannel responses and use back channel responses less frequently.[2]
* Confusion or distraction can occur during an intercultural encounter if participants from both parties are not accustomed to the same backchannel norms.[8] Studies have shown that when people learn a second language they learn or adapt to how people that are native speakers of that language use backchannel responses. This may occur in terms of the frequency that a person produces backchannel responses or what those responses sound like.[2]
* three categories: **non-lexical, phrasal, and substantive**.[11]
  + Non-lexical
    - vocalized sound that has little or no referential meaning but still verbalizes the listener's attention, and that frequently co-occurs with gestures
    - *uh-huh* and *hmm*
    - *In a study examining the use of two-syllable backchannels that focused on mm and mm-hm, Gardner found that the two tokens are generally not identical in function, with mm being used more productively as a continuer, a weak acknowledgment token, and a weak assessment marker. In contrast, mm-hm is generally used as a backchannel to signal that the speaker is yielding their conversational turn and allowing the other speaker to maintain control of the conversational floor.[[13]](https://en.wikipedia.org/wiki/Backchannel_(linguistics)" \l "cite_note-13)*
  + Phrasal / substantive
    - PHRASAL
      * assess or acknowledge a speaker's communication with simple words or phrases (for example, "Really?" or "Wow!"
    - Sustantive
      * Substantive backchannels consist of more substantial [turn-taking](https://en.wikipedia.org/wiki/Turn-Taking) by the listener and usually manifest as asking for clarification or repetitions
    - One of the conversational functions of phrasal backchannels is to assess or appraise a previous utterance. Goodwin argues that this is the case for the phrasal backchannel *oh wow*, where use of the backchannel requires a specific conversational context where something unexpected or surprising was said. Similarly, more substantive backchannels such as *oh come on, are you serious?* require a context where the speaker is responding to something exasperating or frustrating. In both of these cases, Goodwin argues that the backchannels focus only on addressing some aspect of the immediately proceeding utterance rather than the larger conversation itself.[[14]](https://en.wikipedia.org/wiki/Backchannel_(linguistics)" \l "cite_note-14) As a result, they have a broad conversational distribution, appearing both in the middle of extended talk as well as at the end of longer conversational turns.
* **Why to generate backchannels?**
  + Bavelas, Coates, and Johnson[[15]](https://en.wikipedia.org/wiki/Backchannel_(linguistics)" \l "cite_note-:0-15) put forth evidence that listeners' responses help shape the content of the speaker.
  + They concluded that the responses from the distracted listeners included significantly fewer specific responses than from the undistracted listeners. In addition, they found that the quality of the narration was dramatically lower when the listener was distracted. Their basic contention was that listeners are co-narrators and help the storyteller in his or her narration. In other words, a storyteller tells a better story with an audience that is engaged than one that is not.[[15]](https://en.wikipedia.org/wiki/Backchannel_(linguistics)" \l "cite_note-:0-15) **[if listeners distracted]**

**Rapport**

= feeling of being "in sync" with your conversational partner,

=feel the connection and harmony with your partner when you are engaged in a

good conversation

=is argued to underlie many desirable social effects. By generating

proper verbal and nonverbal behaviors, virtual humans have been seen to create

rapport during interactions with human users.

=is argued to underlie successful negotiation

[1], improved quality of child care [2], social engagement [3], and success in teacher-

student interactions [4].

Tickle-Degnen and Rosenberg’s [5] three factor theory of rapport = created through behaviors indicating

* positive emotions (such as head nods or smiles),
* mutual attention (such as mutual gaze),
* coordination (such as postural mimicry or synchronized movements) of nonverbal signals
* as the friendship between two conversants deepens [N/A], the importance of positivity decreases, while the importance of coordination increases
* => **positivity** is more important for our case than coordination/sync !!!

=> Cassell et al. [6] divided rapport into **short-term (building instant rapport)** and long-term (unfolding of both verbal and nonverbal behaviors over the course of a relationship)

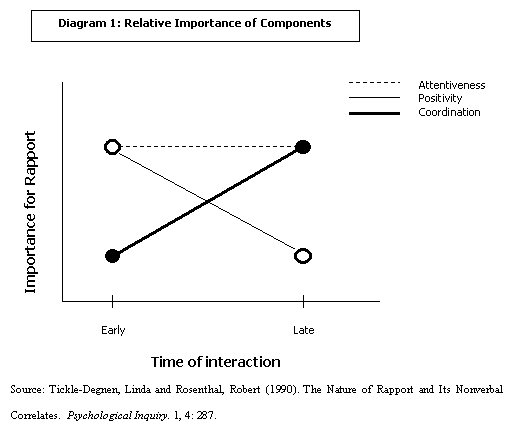
Research on rapport has emphasized the **equal importance of verbal expressions of emotion**, for example, through the reciprocal self-disclosure of hopes and fears [25]

[Wiki - rapport] <https://en.wikipedia.org/wiki/Rapport>

* **Benefits** of interpersonal rapport in domains like education, medicine, or even sales

benefits for psychotherapy and medicine,[6] negotiation,[7] and education,[8]  
smoother interactions, improved collaboration, and improved interpersonal outcomes,[6][7][8] though the specifics differ by the domain.  
(doctor-patient rapport, teacher-student rapport, In negotiation, rapport is beneficial for reaching mutually beneficial outcomes,[[7]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-ReferenceB-7))

* participants engage in reciprocal behaviors such as posture mirroring or in increased coordination in their verbal and nonverbal interactions.[[4]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Psychological_Inquiry-4)
* Rapport building techniques:
  + matching your [body language](https://en.wikipedia.org/wiki/Body_language) (i.e., [posture](https://en.wikipedia.org/wiki/Human_position), [gesture](https://en.wikipedia.org/wiki/Gesture), etc.);[[4]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Psychological_Inquiry-4)
  + indicating attentiveness through maintaining [eye contact](https://en.wikipedia.org/wiki/Eye_contact);[[5]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Intelligent_Virtual_Agents-5)
  + verbal behaviors to increase rapport are
    - the use of positivity (or, positive "face management"),
    - sharing personal information of gradually increasing intimacy (or, "[self-disclosure](https://en.wikipedia.org/wiki/Self-disclosure)"),
    - by referring to shared interests or experiences.[[5]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Intelligent_Virtual_Agents-5)
  + Coordination
    - **Emotional mirroring** (Empathizing with someone's emotional state by being on 'their side'.)
    - **Posture mirroring** (Matching the tone of a person's body language not through direct imitation, as this can appear as mo)
    - **Tone and tempo mirroring** (Matching the tone, tempo, inflection, and volume of a person's voice.)
  + Mutual attentiveness
    - Eye-contact, nodding, smiling
  + Commonality
    - technique of deliberately finding something in common
    - through references to shared interests, dislikes, and experiences.[10]
  + "positive face management",[11]
* Evaluation:
  + Self-reported rapport (on a Likert scale.[[7]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-ReferenceB-7)[[8]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-ReferenceC-8) )
    - most common
    - suffers from unreliability of self-report data, such as the issue of separating participants' reflection on a single interaction with their relationship with the other person more broadly.[12]
  + third-party observer to give a rating of the rapport [[4]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Psychological_Inquiry-4)[[5]](https://en.wikipedia.org/wiki/Rapport" \l "cite_note-Intelligent_Virtual_Agents-5)



**[from Jon Gratch] - howto set window size for rapport generation**

* <http://web.uvic.ca/psyc/bavelas/2000listnrs.pdf>
  + Listeners as co-narrators.
* Prosodic features which cue back-channel responses in English and Japanese
  + => suggests about 1s window size for rapport predictions … but based on speech prosody
  + The fivemost frequent kinds of back-channel feedback were yeah, uh-huh,hm, rightand okay, which iscomparable with what is found in other corpora (Jurafsky et al., 1998) => for paraverbal generation
* Predicting Listener Backchannels:A Probabilistic Multimodal Approach
  + <https://www.researchgate.net/profile/Jonathan_Gratch/publication/225380648_Predicting_Listener_Backchannels_A_Probabilistic_Multimodal_Approach/links/0c9605252ec8bc786e000000.pdf>
  + Rules by Ward and Tsukahara [14] suggest   
    => the informative interval for backchannel predictions is about 1.5s
  + They use window sizes of 0.5, 1, 2 sec with delays: 0, 0.5, 1 s  
    => information from the past as far as
* Lixing dissertation
  + See pdf / email@ict

**[Towards a Dyadic Computational Model of Rapport Management for Human-VirtualAgent Interaction]** <http://articulab.hcii.cs.cmu.edu/wordpress/wp-content/uploads/2015/09/Zhao-Papangelis-Cassell-IVA2014.pdf>

* maintaining rapport between humans and conversational agents over the long-term that operates at the level of the dyad
* explain how humans in dyadic interactions build,maintain, and destroy rapport
* peer tutoring data

[Towards a Computational Architecture of Dyadic Rapport Management for Virtual Agents]

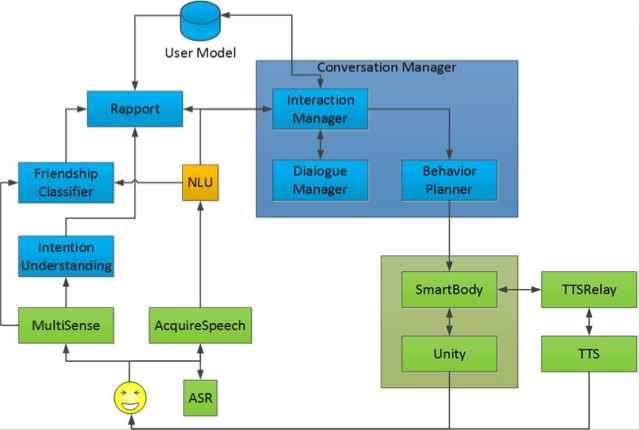
<http://articulab.hcii.cs.cmu.edu/wordpress/wp-content/uploads/2015/09/Papangelis-Zhao-Cassell-IVA2014.pdf>

[http://www.cs.cmu.edu/~rzhao1/I/pub.html#](http://www.cs.cmu.edu/~rzhao1/I/pub.html)

* VA agent architecture - only propose **NOT trained yet**
* Verbal and nonverbal
* **Long-term rapport** agent, complex architecture consisting of many components
* Real-time
* Peer tutoring data
* RL - choose which action
* Inverse RL to estimate reward function

& then direct sparse RL methods to train simulator

* Use MultiSense, SmartBody, Unity,



**[Predicting Levels of Rapport in Dyadic Interactions through Automatic Detection of Posture and Posture Congruence]**

* automatically predicting rapport in dyadic interactions based on posture and congruence - SVMs and MLPs to predict rapport

**[The Nature of Rapport and Its Nonverbal Correlates]** <http://www.justinecassell.com/discourse07/week3/TickleDegnenRosenthal_NatureofRapport.pdf>

* components: mutual attentiveness, positivity, and coor- dination
* in early interactions, positivity and attentiveness are more heavily weighted than coordination, whereas in later interactions, coordination and attentiveness are the more heavily weighted components

**[Rapport with Virtual Agents: What Do Human Social Cues and Personality Explain?]** [**https://ieeexplore.ieee.org/abstract/document/7439777**](https://ieeexplore.ieee.org/abstract/document/7439777)

* investigates how social cues and personality of a human interacting with an agent can be used for automatic prediction of rapport
* Collected: audio-visual data, human personality measures and two measures of rapport: self-reported rapport and rapport judged by observers
* turn-taking patterns and facial expressions are extracted from audio-visual data
* **show** that the most significant cues that infer the rapport judgments are the number of turn-taking cues and pauses
* **confirm** previous findings on how human personality plays an important role in perceiving the interaction with agents-**people who score high in extraversion and agreeableness report higher rapport** with both agents

[Building Rapport between Human and ECA: A Pilot Study] <https://link.springer.com/chapter/10.1007/978-3-319-07230-2_45>

* embodied conversational agents to build rapport with humans
* Paralinguistic behaviors and especially nonverbal beh.
* results suggest that increasing amplitude of nonverbal paralinguistic behaviors may lead to an increased perception of physical connectedness between humans and ECAs.

[Ravaging Resistance:

A Model for Building Rapport in a Collaborative Learning Classroom] <http://radicalpedagogy.icaap.org/content/issue7_1/murphy-valdez.html>

[Customer-Employee Rapport in Service Relationships]

<https://journals.sagepub.com/doi/abs/10.1177/109467050031006>

[Instructor–Student and Student–Student Rapport in the Classroom]

[https://nca.tandfonline.com/doi/abs/10.1080/03634520903564362#.XMdPfUNlD0M](https://nca.tandfonline.com/doi/abs/10.1080/03634520903564362" \l ".XMdPfUNlD0M)

* determine role of rapport in building positive relationships and an overall positive classroom environment

[The Role of Rapport in Investigative Interviewing: A Review]

<https://onlinelibrary.wiley.com/doi/full/10.1002/jip.1386>

**Virtual agents & Rapport**

* VA was more **persuasive** and **better liked** if it mimicked a human speaker's head movement [Bailenson et al. [7]]
* animated agent with text-based dialogue generation that performed nonverbal behaviors such as hand gestures, head nods, eye gaze movements and facial displays of emotion. Their study showed that the agent promotes antipsychotic medication adherence among patients with schizophrenia.
* Rapport Agent could induce the subjective feeling and many of the behavioral benefits of the psychological concept of rapport. - general [12]
* Rapport Agent has proved a valuable tool for advancing Intelligent Virtual Agent (IVA) research, both by demonstrating that virtual agents have important social effects on human users [10,15,16], and by illuminating the factors that contribute towards or sometimes undermine these social consequences [12,15-17].

Current challenges:

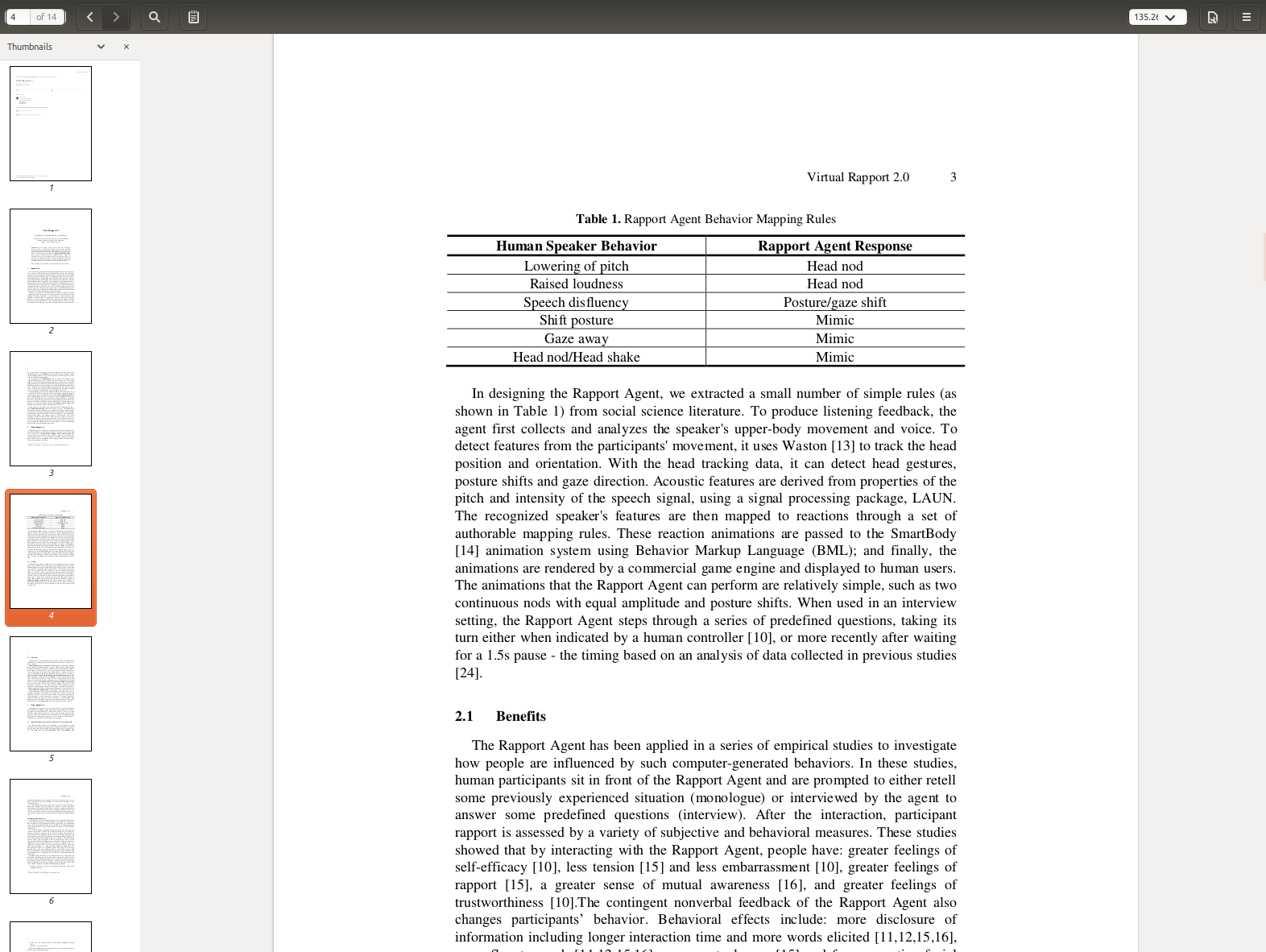
* participants give the system mediocre ratings with respect to subjective measures of rapport and social presence [11]
* generally underperforms human users in terms of subjective and behavioral measures [29] but some subgroups – e.g. shy users – seem to prefer the animated agent [29]
* [31 and THIS] suggest and show that hand-crafted algorithms and animations could be considerably improved

Measures (not very clear how well it is performing at this task)

* timing of backchannel feedback - precision
* Timing of end-of-turn - precision
* performs more natural behaviors

Turn taking

* after waiting for a 1.5s pause - the timing based on an analysis of data collected in previous studies [24].



**VRAs by ICT**

**VRA 1:**

“Creating Rapport with Virtual Agents” [https://ict.usc.edu/pubs/Virtual%20Rapport%202.0.pdf](https://ict.usc.edu/pubs/Virtual Rapport 2.0.pdf)

* Inputs
  + upper-body movement and voice
  + Head position and orientation
    - detect head gestures, posture shifts and gaze direction
  + pitch and intensity of the speech signal
* Outputs
  + SmartBody[14] animation system using Behavior Markup Language (BML)
* **[VRA & effects]** Studies showed that by interacting with the Rapport Agent, people have:
  + greater feelings of self-efficacy [10],
  + less tension [15] and less embarrassment [10],
  + greater feelings of rapport [15], a greater sense of mutual awareness [16],
  + and greater feelings of trustworthiness [10].
  + **[VRA & effects]** Also **changes participants’ behavior**
    - more disclosure of information including longer interaction time and more words elicited [11,12,15,16],
    - more fluent speech [11,12,15,16], more mutual gaze [15] and fewer negative facial expression [17].
* fixed turn-taking strategy
  + Turn when speaker pauses for more than 1.5s

**VRA 2:**

[“Virtual Rapport 2.0” paper / VR2]

= emphasize the importance of data-driven methods for behavior generation

[https://ict.usc.edu/pubs/Virtual%20Rapport%202.0.pdf](https://ict.usc.edu/pubs/Virtual Rapport 2.0.pdf)

* Learn models to predict backchannel and turn-taking opportunity points from the human behavior observed in the same dyadic conversation settings
* parasocial consensus data generates better virtual human behavior than actual human behavior (multiple different views on the same interaction)
* **3 models**
* Backchannel Prediction Model
  + head nods or paraverbals like "um-huh"
  + Predict WHEN and HOW to display the feedback
  + Parasocial consensus data collected using Rapport 06-07 dataset
  + INPUT:
    - **silence T/F**
    - **Eye-gaze T/F**
  + Realtime fwd-only CRF: non-verbal beh. - feedback time
  + 45 videos for training, 4-fold cv
  + OUTPUT: probab of feedback - compare with threshold
  + 3 styles of head nods: track head positions -> cluster -> 3 typical styles (then implement in BML), at runtime find closest ...
  + Randomly choose one
  + Only non-verbal feedback
* End-of-Turn Prediction Model
  + Why? smooth turn-taking strategy without long mutual silence and interruption increases the feeling of coordination
  + Work[19] suggests that **pause** in speech is an **important** **cue** for when to take the turn, but that the amount of time people wait before jumping in can **vary** **considerably** **depending** **on** the **speaker’s nonverbal signals**
  + Multimodal model
    - **silence T/F**
    - **head nods T/F**
    - **eye-gaze T/F**
  + based on the consensus data from the Self-Disclosure data set
  + Rule-based
    - Pause duration, duration of speaker looking at VH, head nod of speaker
  + [say: following the paper VRA2] human speaker is allowed to interrupt VH
    - VH will stop and yield his turn to the human speaker by saying "I'm sorry, keep going" with a regretful facial expression.
* Positive emotion communication / Affective model
  + As described in [VRA2 paper] rapport in initial encounters, can be enhanced by communicating **positive** **emotions** both **nonverbally** and **verbally** (Kang et al. [28])
  + @ backchannel opportunity:
    - If speaker smiles T/F => VH will smile
    - rule-based
  + [28] interviewee discloses more intimate information if the interviewer VH discloses itself first
    - => VH will first disclose info about himself
* **3 stages of the system**
* Perception
  + extracted in real-time are 4 boolean AV features
    - **silence T/F** - Praat, every 100ms
    - **head nods T/F**
    - **eye-gaze T/F** (looking at listener or not)
    - **smile T/F**
* Response models
  + Input: 4 features + VH holds turn or not
    - If **VH holds turn** => **output** from backchannel model **ignored**
* Generation
  + Converted to BML
  + sent to an action scheduler module
  + If **current** **animation** **not** **completed** => **new** animations **ignored**
  + BML passed to the animation system, Smartbody[14]
    - = virtual human animation system designed to seamlessly blend animations and procedural behaviors
  + animations are rendered by a commercial game engine, Gamebryo / Unity, and displayed to users
* Evaluation - subjective [no extensive objective evaluation described, how the models were tuned only mentioned cross-validation]
  + Measures
    - **rapport**, 5-item social presence scales suggested in[26]

(several questions such as "I perceive that I am in the presence of another person in the room with me(1(strongly disagree) -7(strongly agree))")

* **answers** of the five-item social presence scale **highly** **correlated** with each other (the Cronbach's alpha is 0.9) => **average** **them** into one scale
* Significant differences
  + - **overall naturalness,**

(Do you think the virtual agent's overall behavior is natural?(1(not natural at all) -7(absolutely natural)))

* Significant differences
  + - **backchannel feedback** 
      * Precision: How often do you think the virtual human generated feedback at inappropriate time?

(1(all the time) -7(never inappropriate))

* Significant differences
  + - * Recall: How often do you think the virtual human missed feedback opportunities?(1(always miss) -7(never miss)
      * Significant differences
    - **end-of-turn prediction**
      * Correct time: How often do you think the virtual human ask the next question too early?(1(always) -7(never))
      * Significant differences
      * In time: How often do you think the virtual human ask the next question too late?(1(always) -7(never))
      * No significant differences
    - **force-choice task -** which VH do you prefer
  + 2 VHs
  + VH as an interviewer - try to do also as a listener only
  + different question sets for each interaction derived from [25]
  + order of virtual humans and question sets randomized
  + 21 participants - evaluated **both** VHs
  + Evaluate significant differences along the 4 measures between VRA 1 and VRA 2
  + Results: VRA 2:
    - more "in sync" with the human speaker during the interaction
    - Strengthens reciprocity by self-disclosure
    - predictsthe timing of backchannel feedback and end-of-turnmore precisely,
    - performs more natural behaviors and thereby creates much stronger feelings of rapport between users and virtual agents
* Future work
  + plan to deploy ourvirtual human in various scenarios to assess how it will influence the human partnerin different situations
* From audio signal they extract only 1 boolean feature (speaker is silent or not)

**Turn-taking**

* take a turn (ask the next question) too quickly - most likely associated with negative and strong personality [23]   
  (opposite to the goal of creating rapport)

**Head nod/shake/tilt detection**

* There is a low to moderate correlation(Pearson’s r=0.42) (Dancey and Reidy2004) between number of words spokenand number of head movements produced, which explains 58 % of the variation. Inturn, this means that there must also be genuine individual variation in how muchspeakers use head movements when they speak

Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions

* <https://iui.ku.edu.tr/wp-content/uploads/2018/06/is2018_cameraReady.pdf>
* Predict head nod & trun-taking
* For realtime
* SVM (with feature summarization) RBF kernel, LSTM 50 units, one layer
* 5-fold cv
* Human-human conversational data
* IEMOCAP dataset
  + They made annotations: nod and gaze, laugh/smile(previous paper)
  + Turn taking - was already there
* Audio-visual
* Audio features
  + Spectral 13 MFCC
  + Prosody 6 dimensions
  + Z norm over temporal window
* Video features
  + 3D binary: nod, smile/laugh, gaze away
* Sync at 40 Hz
* POI (make predictions on) - OTH
* Like virtual human - human
* Turn taking prediction
  + AV of VH + V of H (if VH speaking)
  + V of VH + AV of H (if H speaking)
  + 2 sec window
* Nod prediction
  + AV of VH + AV of H
  + **We only want AV**
  + 3 sec window
* Predict different #frames in future
* Decision fusion of SVM and LSTM

**Generative models**

* Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration
  + <https://arxiv.org/abs/1702.08212>
  + CVAE model
    - Fully-connected 5 layers, 2 latent variables z\_t-1 and z\_t
    - N past frames (flattened) -> N future frames
    - No recurrent connections/extensions
    - Train separate models for different tasks (main body, left, right)
  + Max 1660ms in future, but 300-500ms optimal
  + Average prediction error - MSE over all windows, also report variance
  + Sampling at encoder and transitioner stage: z\_t-1 and z\_t