CHAPTER 14

MODELING HUMAN COMMUNICATION DYNAMICS FOR VIRTUAL HUMAN

Louis-Philippe Morency, Ari Shapiro and Stacy Marsella
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA
morency@ict.usc.edu, shapiro@ict.usc.edu, marsella@ict.usc.edu

1. Introduction

Face-to-face communication is a highly interactive process where participants mutually exchange and interpret linguistic and gestural signals. Communication dynamics represent the temporal relationship between these communicative signals. Even when only one person speaks at a time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. The transactional view of human communication shows an important dynamic between communicative behaviors where each person serves simultaneously as speaker and listener (Watzlawick et al., 1967). At the same time you send a message, you also receive messages from your own communications (individual dynamics) as well as from the reactions of the other person(s) (interpersonal dynamics) (DeVito, 2008).

Individual and interpersonal dynamics play a key role when a teacher automatically adjusts his/her explanations based on the student nonverbal behaviors, when a doctor diagnoses a social disorder such as autism, or when a negotiator detects deception in the opposite team. An important challenge for artificial intelligence researchers in the 21st century is in creating socially intelligent robots and computers, able to recognize, predict and analyze verbal and nonverbal dynamics during face-to-face communication. This will not only open up new avenues for human-computer interactions but create new computational tools for social and behavior researchers–software able to automatically analyze human social and nonverbal behaviors, and extract important interaction patterns.

Human face-to-face communication is a little like a dance, in that participants continuously adjust their behaviors based on verbal and nonverbal behaviors from other participants. We identify four important types of dynamics during social interactions:

- **Behavioral dynamic** A first relevant dynamic in human

communication is the dynamic of each specific behavior. For example, a smile has its own dynamic in the sense that the speed of the onset and offset can change its meaning (e.g., fake smile versus real smile). This is also true about words when pronounced to emphasize their importance. The behavioral dynamic needs to be correctly represented when modeling social interactions.

- **Multimodal dynamic** Even when observing participants individually, the interpretation of their behaviors is a multimodal problem in that both verbal and nonverbal messages are necessary to a complete understanding of human behaviors. Multimodal dynamics represent this influence and relationship between the different channels of information such as language, prosody and gestures. Modeling the multimodal dynamics is challenging since gestures may not always be synchronized with speech and the communicative signals may have different granularity (e.g., linguistic signals are interpreted at the word level while prosodic information varies much faster).
- **Interpersonal dynamic** The verbal and nonverbal messages from one participant are better interpreted when put into context with the concurrent and previous messages from other participants. For example, a smile may be interpreted as an acknowledgement if the speaker just looked back at the listener and paused while it could be interpreted as a signal of empathy if the speaker just confessed something personal. Interpersonal dynamics represent this influence and relationship between multiple sources (e.g. participants). This dynamic is referred as micro-dynamic by sociologists (Hawley, 1950).
- **Societal dynamic** We categorize the organizational (often referred as meso-level) and societal (often referred as macro-level) dynamics in this general category which emphasize the cultural change in a large group or society. While this proposal does not focus on societal dynamics, it is important to point out the bottom-up and top-down influences. The bottom-up approach emphasizes the influence of micro-dynamics (behavioral, individual and interpersonal) on large-scale societal behaviors (e.g., organizational behavior analysis based on audio micro-dynamics (Pentland, 2004)). As important is the top-down influence of society and culture on individual and interpersonal dynamics.

In this book chapter, we first discuss techniques to model the behavior dynamics of virtual human as well as human participants. We then address the challenge of multimodal dynamic and, more specifically, the synchrony between speech and gestures for virtual human. Finally, we present approaches to model the interpersonal dynamic between speaker and listeners using state-of-the-art machine learning. We finally conclude by

discussing the future challenges related to societal and cultural dynamics.

2. Behavior Dynamic: Virtual Human Animation

Virtual humans can be used to express portray a wide range of behaviors, including synchronized speech, gestures and facial expressions. In order to generate such expressions on a virtual human, an animation must be synthesized and replayed on a digital character. Such an animation can be generated from a variety of sources, such as motion capture, hand-designed by a digital artist, or procedurally generated from a motion synthesis algorithm. The synthesized motion can be coordinated with audio and lip motion on the digital character if the gesture is associated with an utterance. Digital 3D characters that are human-like in appearance and responses are termed virtual humans.

The use of digital characters is common in video games, simulations and live action and animated feature films. The most widely used method of generating a 3D character in motion involves combining a 3D geometric mesh, its surface colors or images, called textures, with a set of joints combined in a hierarchy, called a skeleton. The skeleton is then used to modify the 3D geometry via a deformation, or skinning process, where each joint of the skeleton is related to one of more of the faces of the geometry and modifies the geometry as the position and orientation of the joint changes. Thus, 3D characters that consist of a 3D mesh can move, stretch and twist such a mesh in a way that appears somewhat natural by simply rotating and translating the joints in the underlying skeleton. Thus, it is often sufficient to generate the motion for a 3D character in order to appropriately express a gesture or movement when using these common 3D character animation techniques. Thus the challenge of synthesizing gestures is generally associated with acquiring and synthesizing skeletal motion.

Other methods of generating digital characters and motions exist, such as the use of image-based techniques, where an entire mesh, texture and movement are captured at the same time. However, at the time of this writing, such methods are not mature and are not widely supported.

There are several different kinds of gesture and expression architectures that can be used, each offering a different level of quality, different data requirements, different flexibility and varying complexity of use. In general, the gesture architectures that offer the highest level of quality are those that use motion capture data explicitly and replay it without modification. Those with the lowest level of quality segment gestures into smaller phases then synthesize motions procedurally from various algorithms. Gesture architectures with the greatest level of control and precision generally favor the reverse; procedurally generated, phased gesture motion provides better control than replayed motion capture clips. The following sections describe variations in a gesture architecture based on these ideas.

Figure 1: An actor using an LED-based motion capture suit [PhaseSpace]. In this session, the body motion is being captured, but not the facial performance or audio/utterance.

Figure 2: A hub-and-spoke gesture architecture. An underlying idle pose is created (center character), from which a set of gestures can be played. This allows individual gestures to be used for different utterences. Note that a different set of gestures must be generated for each underlying pose, which can vary from standing, to sitting, to standing with various hand poses and body lean.

2.1 Motion Capture Session

In general, a motion capture session requires a motion capture system based on cameras or inertial sensors. The session typically requires

transferring the data onto a virtual character during the capture session. The motion capture process typically synthesizes data onto an existing skeleton, which does not match exactly the proportions and sizes of the real human actor. Thus, the captured data needs to be retargeted or transferred onto the skeleton which models the virtual character. Motion capture data is typically segmented into clips, each clip representing semantically-related content. Similarly, longer motion capture clips can be segmented into smaller ones during a postprocessing phase, where data is refined and edited. The specifics of capturing motion via motion capture can be found in other sources. Below we describe several capture and motion synthesis strategies.

2.2 Full Performance Architectures

The highest level of quality can be obtained by simultaneously capturing both the utterance and the gesture. Thus, a performer will act out an utterance in combination with its associated body movements, including gestures. This performance will then be replayed in its entirety on a virtual human. The advantage to this approach is that the virtual human will be able to faithfully replay the human performance, notwithstanding the retargeting necessary to fit the captured performance onto the virtual human's skeleton. However, one of the difficulties of using this method is that the audio of the utterance in combination with the associated facial movements must be synchronized with the body animation. In other words, the body movement, facial movement and audio track must be synchronized together exactly as they were captured. This means capturing the facial animation performance at the same time as the body movement (Stone et al 2004), which typically requires three different capture systems; a motion capture system for the body, a separate one for the face, and a third audio capture for the utterance. In order to simplify this process, the facial performance can be captured during a separate performance, but doing so risks a lack of synchronization with the original body performance. While this methods achieves the highest possible level of quality when using the standard 3D asset pipeline, they also require a great amount of effort in order to manually synchronize the three main components together. In addition, the high level of quality comes at the expense of specificity; the performance is only meaningful in contexts similar to those during the recorded session. For example, consider an actor whose performance in a dialogue is captured and synthesized onto a virtual character. By only capturing one of the two actors and synthesizing that performance onto a virtual human, you risk misapplying the subtleties of the actor during the performance that are in response to the presence or movements of the other actor. A recording interaction might be subtlety different when used in a different conversation with a different person. Conversational energy,

timing, backchannelling and even gaze can be different with different partners. Thus the high level of quality achieved by replaying an actor's performance can be limited in its use outside of the original context.

2.3 Hub-and-Spoke Architectures

As an alternative to reusing a motion captured performance directly, a hub-and-spoke architecture can be used to achieve greater reuse of speech, gestures and facial performance. This method uses an underlying base, or idle pose, and blends gesture motions that start and end in a similar position as the idle pose (Shapiro 2011). For example, an actor will perform a number of gestures starting from the base pose, performing the gesture, and then returning to the base pose. Thus each gesture can be replayed on a virtual human in different order with other gestures, each starting and ending from the same base pose, which is usually implemented as a continuous idle posture. This method allows you to synthesize an arbitrary sequence of gestures, while maintaining a high level of motion quality, since each individual gesture maintains the nuances of the original performance. A drawback of this method is the large number of gesture performances needed for each base pose or posture. For example, a different performance would be needed for gestures when standing up with your hands relaxed at your side, versus standing up with your hands on your hips, versus standing up with only one hand resting on your hip, and so forth. Posture changes, such as sitting down, crossing your legs and so forth, would also require a new set of gestures for each posture. Note that speech and facial movements can be simultaneously recorded with the gestures when using the hub-and-spoke method, or synthesized at a later time. Also note that using such a method limits the types of performances that can be captured; the actor must return to the same position that he started from, perhaps causing a lack of continuity between utterances. However, this approach's strength is in its ability to sequence gestures in any order or time as needed.

2.4 Blending Gestures to Achieve Motion Variation

While the best quality reproduction will be achieved by replaying a performance exactly as it were captured, it is often impractical or unfeasible to capture all the movements that would be replayed on a virtual human. For example, a pointing gesture can take many forms; accusatory, informational, subtle and so forth. Thus, variations in gesture can be generated by multiple performances of an actor of those gesture variations. This will result in a high-quality reproduction of the gesture. This also means that additional motion captured resources will be needed to capture the additional gesture variation, and itself will only be reusable for one additional synthesized performance that matches the new variation. To remedy this problem, similar gesture motions can be blended together in order to give a range of variation between various example motions. For

Figure 3: A simplified hierarchy for synthesizing motion. The rectangles indicate the parts of the body controlled during each state, starting from the larger rectangles and going inward towards the smaller rectangles. The entire body motion is synthesized for, in this case, a sitting virtual human. Next, the upper body gesturing is synthesized by layering gesture movements on top of the lower body. Next, spine movements controlling posture and gaze are layered on top of the gesture. Then, head movements are added for backchannelling and movement during speaking. Next, facial movements to express emotion and coordinate lip movement, then eye movement including saccades and eyelid positioning.

example, an energetic beat gesture can be combined with a slow, deliberate beat gesture to generate a gesture motion that appears halfway between an energetic and a slow gesture. Likewise, directional gestures, such as pointing gestures, can be generated with directional variations, then recombined in order to form a pointing gesture in a direction not explicitly captured, but rather synthesized as a combination of two or more other pointing gestures. There are several limitations to this approach; gestures that are to be blended together need to have compatible characteristics in order for the blending to work properly. In most cases, this means that the motions much have the same number and type of phases. One motion cannot, for example, have three small shakes of the hand, while the other only has two. In addition, large movements across the body or broad variations in poses across the gestures blend together poorly. Blended poses can vary in completion time; it is unlikely that any two motion captured gestures will take the same amount of time. Gestures are blended together by first timewarping, that is stretching or compressing, the motions to the desired time, and then combining the various motions together. A large difference in time between any two motions will lead to poor quality blends, since one or both motions will need to be lengthened or shortened to match the other, typically changing the dynamics of motion that are embedded within the original captured motion. Thus a gesture that is

synthesized from one or more blends maintains the highest level of fidelity when the individual blended gestures have matching phases, and similar timings. There are many different ways to blend motions together, offering various trade-offs between execution time and memory (Kovar and Gleischer 2004, Huang and Kallmann 2010) as well as tradeoffs between precision and smoothness (Pettre et al 2006, Rose et al 2001). An overview of common blending techniques can be found in Feng et al (2012).

2.5 Hierarchical Gesture Models

One model for achieving variations in gestures is to use a model of hierarchical of control over the virtual human movement (Kallmann and Marsella 2005). In this model, virtual human movement is divided into a generalization/specialization hierarchy. Thus, movement is first performed for the entire body, usually a sitting or standing pose, then a gesture movement using the arms and torso is performed, then a separate head and neck movement, then facial and eye movement. By layering such animations together, it becomes possible to achieve a large variation in gesture performance. For example, the same gesture can be combined with several different head or face movements, producing differing performances. In addition, the hierarchical nature allows the integration of procedural elements such as gaze control (Thiebaux et al., 2008) to override specific parts of the body in order to modify the underlying motion for the specific context in which the motion in used. The drawback to using such architecture is the loss of fidelity of the resulting motion; since the synthesized motion was never originally captured on a human, the dynamics and subtleties of the synthesized motion will differ from that of a human actor performing the same motion. Thus, using a hierarchy to generate gesture performance yields a large variation in performance, at the expense of motion quality.

3. Multimodal Dynamic: Speech and Gestures Generation

The generation of multimodal behavior for a virtual human faces a range of challenges. Must fundamentally is the question of what behaviors to exhibit. Nonverbal behaviors serve a wide variety of communicative functions in face-to-face interaction. They can regulate the interaction: a speaker may avert gaze to hold onto the dialog turn and may hand-off the turn by gazing at a listener. The speaker can use nonverbal behaviors to convey propositional content: a nod can convey agreement, raising eyebrows can emphasize a word. The propositional content of the nonverbal behavior can stand in different relations to the verbal content, providing information that embellishes, substitutes for and even contradicts the information provided verbally. In other words, the nonverbal behavior is not simply an illustrator

of the verbal information. Nonverbal behaviors also convey a wide range of mental states and traits: gaze aversions can signal increased cognitive load, blushing suggest shyness and facial expressions can reveal emotional states.

Another challenge here is that this mapping between communicative function and behaviors is many-to-many. One can emphasize aspects of the dialog using a hand gesture, a nod or eyebrow raise. On the other hand, a nod can be used for affirmation, emphasis or to hand over the dialog turn. The context in which the behavior occurs can transform the interpretation, as can subtle changes in the dynamics of the behavior: head nods signaling affirmation versus emphasis typically have different dynamics. Further, behaviors can be composed with each other, further transforming their interpretation.

Additionally, the behaviors are often tightly synchronized and changes in this synchronization can lead to significant changes in what is conveyed to a listener. For instance, the stroke of a hand gesture, a nod or an eyebrow raise individually or together are often used to emphasize the significance of a word or phrase in the speech. To achieve that emphasis the behavior must be closely synchronized with the utterance of the associated words being emphasized. Alteration of the timing will change what words are being emphasized and consequently change what is conveyed to a listener.

Achieving such synchronization in a virtual human can be difficult, especially in the case of behaviors such as hand gestures that involve relatively large-scale motion and multiple phases. Consider a beat gesture, a staccato, often downward stroke of the hand that can be used to provide emphasis. To perform a downward motion, the hand must be raised in preparation for the stroke. After the stroke, the hand can be held in a pose to provide further emphasis, followed by a relaxation to a rest position. This sequence of behaviors occur in alignment with the speech, so there must be sufficient time to prepare for the stroke, the stroke and any post-stroke hold must be tightly coordinated with the parts of the dialog that is being emphasized. Further the relaxation may need to take into account co-articulation, that there will be subsequent gestures to be performed.

In addition to this synchronization between the speaker's behaviors, there is also the issue of synchronization between speaker and listeners as the speaker's utterance unfolds. Listeners exhibit a variety of behaviors both generic feedback that signals the listener is attending to the speaker and that the speaker should continue as well as specific feedback tied to a deeper understanding of, and cognitive/emotional reactions to, the personal relevance of what the speaker is saying as the utterance unfolds (Bavelas et

al.,). The speaker can in turn dynamically adapt what they are saying in response to this feedback.

Such dynamic adjustments raise challenges in generating both the verbal and nonverbal behaviors dynamically and incrementally for a virtual human. Ideally a virtual human listener should respond to a human speaker, providing generic feedback signaling attention as well as more specific feedback that signals comprehension and reaction to the speaker's unfolding utterance. This requires a natural language system that can incrementally understand the human speaker's utterance. Conversely, the virtual human as it is speaking should be aware of a human listener's behavior, responding to nonverbal signals such as confusion. Together such dynamic interaction suggests capabilities such as interruption of behavior as well as incremental understanding and generation of verbal and nonverbal behavior.

3.1 An Approach to Nonverbal Behavior Generation

A range of systems have tackled various aspects of these challenges (refs). Here we discuss one of the approaches: The Nonverbal Behavior Generator (NVBG) (Lee & Marsella, 2006; Wang et al. 2011). NVBG automates the generation of physical behaviors for virtual humans, including nonverbal behaviors accompanying the virtual humans dialog, responses to perceptual events as well as listening behaviors. It takes input from the virtual human's knowledge of its task, dialog, emotional reactions to events and perceptual processes. Modular processing pipelines transform the input into behavior schedules, written in the Behavior Markup Language (BML, Kopp et al., 2006) and then passed to a character animation system (SmartBody, Thiebaux et al., 2008).

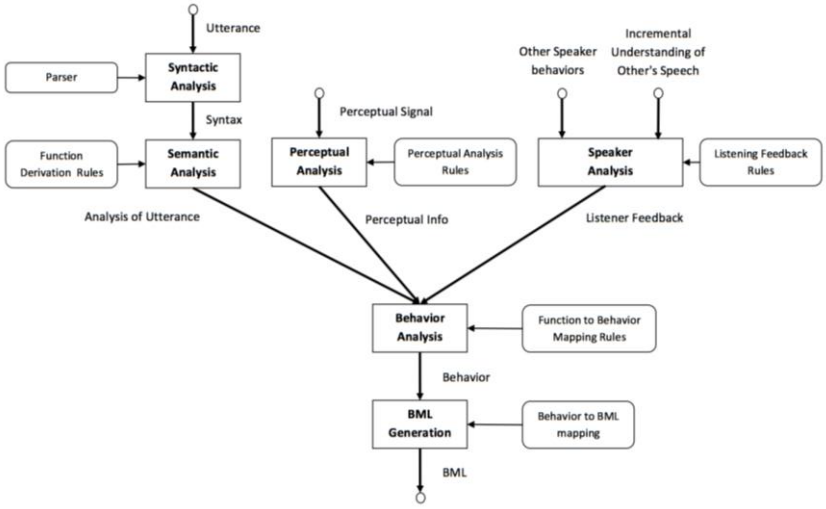

Figure 4: Overview architecture for verbal and nonverbal behavior generation in a virtual human.

Error! Reference source not found. depicts three pipelines used to generate behavior for the virtual human's including co-verbal non-verbal behavior, reaction to events and listening behavior. The sections below discuss these three pipelines. Note the initial processing differs but eventually merges at the behavior analysis.

3.2 Processing the virtual human's utterances

The utterance pipeline (left side in Figure 4) analyzes the surface text of the virtual human's utterance to infer appropriate nonverbal behavior. The processing in NVBG does not make any strong assumption about the input's markup of the agent's communicative intent or internal state (e.g. affective state, attitude). When such information is missing, the system attempts to infer it (essentially falling back on the more limited role of illustrating and embellishing the language channel) For instance, in the absence of detailed markup of the virtual human's communicative intent, such as points of emphasis or emotion, NVBG analyzes the surface text to support the generation of believable nonverbal behaviors.

To this end, the sentence is first parsed to derive the syntactic structure. Then a semantic analysis phase attempts to infer aspects of the utterance's communicative function using inference rules to build up a hierarchical structured lexical, semantic and pragmatic analysis. Examples of these communicative functions include affirmation, inclusivity, intensification, etc. (see [Lee & Marsella, 2006] for details). NVBG then goes through a

behavior analysis stage, in which a set of *nonverbal behavior rules* map from communicative functions to classes of nonverbal behaviors. A BML generation phase then maps those behavior classes to specific behaviors, described in BML. This mapping can use character specific mappings designed to support differences including personality, culture, gender and body types. Conflict resolution occurs at several phases in the overall process. For example, if there are two or more rules overlapping with each other causing conflict, NVBG resolves the conflict by filtering out the rule with lower priority. The priority value of rules has been set through a study of human behaviors using video corpora. The final result is a schedule of behaviors that is passed to the character animation system.

Research on NVBG has explored several approaches to encoding the knowledge used in the function derivation and behavior mapping. Initial work on NVBG was based on an extensive literature review of the research on nonverbal behavior. This seeded the development of rules encoding the function derivation and behavior mapping rules. Then videos of real human face-to-face interactions were annotated and analyzed to verify the rule knowledge, embellish knowledge with dynamic information about behaviors and develop a conflict resolution system that is used to resolve conflicts between behavior suggestions. This annotation and analysis was critical because existing literature said little about dynamics of behaviors and further conflict resolution was to resolve potential conflicts both between the behaviors suggested by the rules as well as differences across literature sources.

More recently a variety of machine learning techniques have been explored, including Hidden Markov Models and Latent-Dynamic Conditional Random Fields to learn the mapping between features of an utterance and nonverbal behaviors using annotated data face-to-face interactions. In particular, Lee & Marsella (2012) contrasts several approaches to learning models of head and eyebrow movement as well as contrasting the results with the knowledge encoded in NVBG by the literature approach discussed above.

3.3 Perceptual Processing

Perceptual messages are treated differently than generating nonverbal behavior for the virtual human's utterances. For the perceptual messages, NVBG is deciding on how to respond to signals about external events, including the physical behavior of objects, humans or other virtual humans. These responses, such as looking at a moving object, can in large measure be reflexive or automatic as opposed to having an explicit communicative intention like an utterance. Due to the differences between the perceptual

and utterance use cases, NVBG's perceptual responses analyses use a different processing pipeline than the utterance handling.

Specifically, NVBG's response is determined by a Perceptual Analysis stage that leads into the Behavior Analysis and BML Generation stages discussed previously. The rules used during Perceptual Analysis take into account what is the perceived behavior and whether the perceived behavior is above some acceptance threshold (e.g., an object's speed, size and distance or an event's duration or magnitude).

3.4 Listener Feedback

The listener feedback pipeline handles the virtual human's behavior while listening to a human or virtual human speaker. The approach makes a distinction between generic feedback and specific feedback, handling them using different rule sets. Generic feedback is driven by speaker behaviors including nods and pauses in speech. Specific feedback is driven by the virtual human's unfolding interpretation of, and reaction to, the speaker's utterance, which requires natural language technology that provides incremental interpretation of partial utterances such as the work of Devault et al., 2011 which provides a semantic interpretation, a measure of confidence in the current understanding and a measure of whether continued listening will lead to better understanding. The virtual human's reaction to the understanding is a valenced reaction to the evolving interpretation of the speaker's utterance. For example, if the virtual human's interprets the speaker's partial utterance as deliberately proposing an action to harm the virtual human, then the reaction will be anger.

The model analyzes this information and triggers relevant listener feedback rules, which are mapped to appropriate nonverbal behaviors, such as nods for generic feedback and expressions of confusion, comprehension, happiness or anger for the specific feedback. These behaviors are also conditional on the listener's roles and goals. In particular, a listener that is the main addressee and hos the goals of participating in and understanding the conversation will engage in mutual gaze with the speaker, nod to signal attention and signal comprehension and reaction to the content of the utterance. On the other hand an eavesdropper that has the goal of avoiding participation in the conversation will avoid mutual gaze and signaling attention with nods.

4. Interpersonal Dynamic: Speaker and Listener Interaction

A great example of interpersonal dynamics is backchannel feedback, the

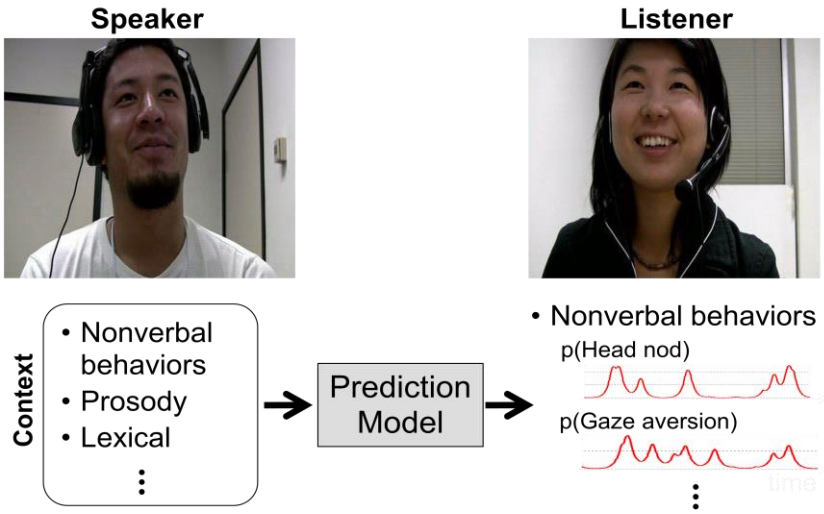

Figure 5: Contextual prediction example: online prediction of the listener's backchannel based on the speaker's contextual features. In the contextual prediction framework, the prediction model automatically (1) learns which subset of the speaker's verbal and nonverbal actions influences the listener's nonverbal behaviors, (2) finds the optimal way to dynamically integrate the relevant speaker actions and (3) outputs probabilistic measurements describing how likely listener nonverbal behavior are.

nods and para-verbals such as "uh-huh" and "mm-hmm" that listeners produce as someone is speaking (Watzlawick et al., 1967). They can express a certain degree of connection between listener and speaker (e.g., rapport), a way to show acknowledgement (e.g., grounding) or they can also be used for signifying agreement. Backchannel feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participant's ability to communicate (Bavelas et al. 2000). Accurately recognizing the backchannel feedback from one individual is challenging since these conversational cues are subtle and vary between people. Learning how to predict backchannel feedback is a key research problem for building immersive virtual human and robots. Finally, there are still some unanswered questions in linguistic, psychology and sociology on what triggers backchannel feedback and how it differs from different cultures. In this chapter we show the importance of modeling both the multimodal and interpersonal dynamics of backchannel feedback for recognition, prediction and analysis.

Description of the approach to model listener backchannel feedback based on speaker behaviors (see Figure 5). We first describe the main ideas behind contextual prediction, then show some of the audio-visual features

commonly used and finally describe recent prediction models.

4.1 Contextual Prediction

Natural conversation is fluid and highly interactive. Participants seem tightly enmeshed in something like a dance, rapidly detecting and responding, not only to each other's words, but to speech prosody, gesture, gaze, posture, and facial expression movements. These ``extra-linguistic'' signals play a powerful role in determining the nature of social exchange. When these signals are positive, coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in such diverse areas as negotiations and conflict resolution (Drolet & Morris, 2000; Goldberg, 2005), psychotherapeutic effectiveness (Tsui, 2005), improved test performance in classrooms (Fuchs, 1987), and improved quality of child care (Burns, 1984). Not surprisingly, supporting such fluid interactions has become an important topic of human-centered research.

In the contextual prediction framework, the prediction model automatically learns which subset of a speaker's verbal and nonverbal actions influences the listener's nonverbal behaviors, finds the optimal way to dynamically integrate the relevant speaker actions and outputs probabilistic measurements describing the likelihood of a listener nonverbal behavior. Figure 5 presents an example of contextual prediction for the listener's backchannel.

The goal of a prediction model is to create online predictions of human nonverbal behaviors based on external contextual information. The prediction model learns automatically which contextual feature is important and how it affects the timing of nonverbal behaviors. This goal is achieved by using a machine learning approach wherein a sequential probabilistic model is trained using a database of human interactions. A sequential probabilistic model takes as input a sequence of observation features (e.g., the speaker's features) and returns a sequence of probabilities (e.g., of the listener's backchannel). Some of the most popular sequential models are the Hidden Markov Model (HMM) (Rabiner, 1989) and the Conditional Random Field (CRF) (Lafferty et al., 2001). A main difference between these two models is that the CRF is discriminative (i.e., tries to find the best way to differentiate cases where the human/agent produces a particular behavior or does not) while the HMM is generative (i.e., tries to find the best way to generalize the samples from the cases where the human/agent produces a behavior without considering the cases where no such behavior occurs). The contextual prediction framework is designed to work with any types of sequential probabilistic models.

At the core of the approach is the idea of context, the set of external factors which can potentially influence a person's nonverbal behavior.

4.2 Context (shallow features)

Conceptually, context includes all verbal and nonverbal behaviors

performed by other participants (human, robot, computer or virtual human) as well as the description of the interaction environment. For a dyadic interaction between two humans (as shown in Figure 5), to predict the nonverbal behavior of the listener, the context will include the speaker's verbal and nonverbal behaviors, including head movements, eye gaze, pauses and prosodic features. What differentiates the computation framework from "conventional" multi-modal approaches (e.g., audio-visual speech recognition) is that the influence of other participants (and the environment) on a person's nonverbal behavior is modeled directly instead of only modeling signals from the same source (e.g., the listener in Figure 5).

In previous work, four types of contextual features were highlighted: lexical features, prosody and punctuation features, timing information and gesture displays. Such features were used to recognize human nonverbal gestures, when the robot spoke to a human, or to generate a gesture for a virtual human given a human's verbal and nonverbal contributions in an interaction (Morency et al., 2008).

Shallow versions of each of these features were calculated either automatically or manually annotated from the dialogue manager of the robot (or virtual human) or directly from a human's action. For lexical features, individual words (unigrams) and word pairs (bigrams) provided information regarding the likelihood of gestural reaction. A range of techniques were used for prosodic features. Using Aizula system (Ward and Tsukahara, 2000), pitch, intensity and other prosodic features were automatically computed from the human's speech (Morency et al., 2008). With robots and virtual humans, the generated punctuation was used to approximate prosodic cues, such as pauses and interrogative utterances. Synthesized visual gestures are a key capability of robots and virtual humans, and they can also be leveraged as a context cue for gesture interpretation. The gestures expressed by the ECA influences the type of visual feedback from the human participant. For example, if the agent makes a deictic pointing gesture, the user is more likely to look at the location that the ECA is pointing to; in human-human dialogues, a critical gestural feature was where the speaker looked. This demonstrates that shallow, very simple features are reliably useful in predicting nonverbal gestures.

The shallow features used in previous work were easy to calculate or annotate and were used for both ECA-human and human-human interactions. However, the features' very simplicity allows them to capture only a small part of the information available in linguistic and gestural behavior.

4.3 Modeling Latent Dynamic

One of the key challenges with modeling the individual and interpersonal dynamics is to automatically learn the synchrony and

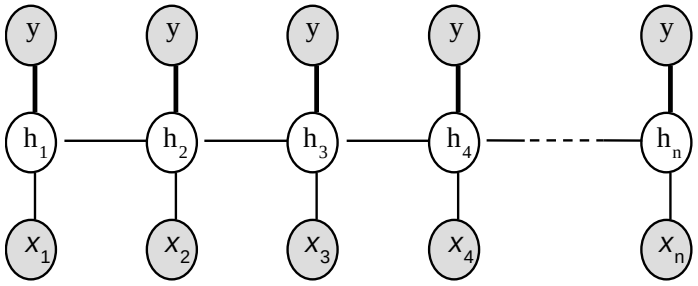

Figure 6: Graphical representation of the LDCRF model. x_j represents the j^{th} observation (corresponding to the j^{th} observation of the sequence), h_j is a hidden state assigned to x_j , and y_j the class label of x_j (i.e. positive or negative). Gray circles are observed variables.

complementarities in a person's verbal and nonverbal behaviors and between people. A new computational model called Latent-Dynamic CRF (see Figure 6) was developed to incorporate hidden state variables that model the sub-structure of a class sequence and learn dynamics between class labels (Morency et al., 2007). It is a significant change from previous approaches which only examined individual modalities, ignoring the synergy between speech and gestures.

The task of the Latent-Dynamic CRF model is to learn a mapping between a sequence of observations $\boldsymbol{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\boldsymbol{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set Y of possible class labels, for example, Y = {*backchannel*, *other-gesture*}. Each observation x_j is represented by a feature vector $\phi(x_j)$ in R^d, for example, the head velocities at each frame. For each sequence, a vector of ``sub-structure'' variables $\boldsymbol{h} = \{h_1, h_2, ..., h_m\}$ is assumed. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Given the above definitions, the latent conditional model is defined as:

$$P(\boldsymbol{y} \mid \boldsymbol{x}, \theta) = \sum_{\boldsymbol{h}} P(\boldsymbol{y} \mid \boldsymbol{h}, \boldsymbol{x}, \theta) P(\boldsymbol{h} \mid \boldsymbol{x}, \theta)$$

where θ are the parameters of the Latent-Dynamic CRF model. These are learned automatically during training using a gradient ascent approach to search for the optimal parameter values. More details can be found in Morency et al. (2007).

The Latent-Dynamic CRF model was applied to the problem of learning individual dynamics of backchannel feedback. Figure 7 shows the LDCRF model compared previous approaches for probabilistic sequence labeling (e.g. Hidden Markov Model and Support Vector Machine). By modeling the hidden dynamic, the Latent-Dynamic model outperforms previous approaches. The software was made available online on an open-source website (sourceforge.net/projects/hcrf).

4.4 Prediction Model of Interpersonal Dynamics

In the contextual prediction framework, the prediction model automatically learns which subset of a speaker's verbal and nonverbal actions influences the listener's nonverbal behaviors, finds the optimal way to dynamically integrate the relevant speaker actions and outputs probabilistic measurements describing the likelihood of a listener nonverbal behavior. Figure 5 presents an example of contextual prediction for the listener's backchannel.

The goal of a prediction model is to create online predictions of human nonverbal behaviors based on external contextual information. The prediction model learns automatically which contextual feature is important and how it affects the timing of nonverbal behaviors. This goal is achieved by using a machine learning approach wherein a sequential probabilistic model is trained using a database of human interactions.

The contextual prediction framework can learn to predict and generate dyadic conversational behavior from multimodal conversational data, and applied it to listener backchannel feedback (Morency et al., 2008). Generating appropriate backchannels is a notoriously difficult problem because they happen rapidly, in the midst of speech, and seem elicited by a variety of speaker verbal, prosodic and nonverbal cues. Unlike prior approaches that use a single modality (e.g., speech), it incorporated multimodal

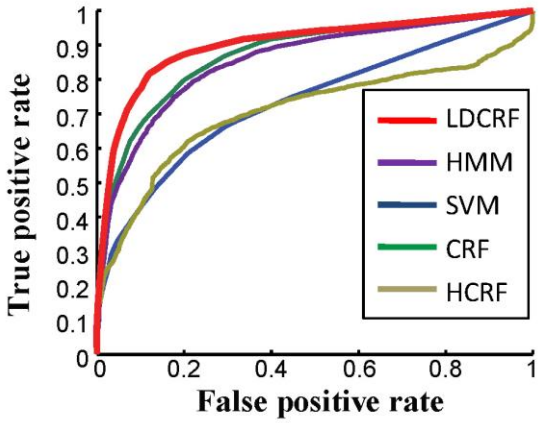

Figure 7: Recognition of backchannel feedback based on individual dynamics only. Comparison of the Latent-Dynamic CRF model with previous approaches for probabilistic sequential modeling.

features (e.g., speech and gesture) and devised a machine learning method that automatically selects appropriate features from multimodal data and produces sequential probabilistic models with greater predictive accuracy

4.5 Signal Punctuation and Encoding Dictionary

While human communication is a continuous process, people naturally segment these continuous streams in small pieces when describing a social interaction. This tendency to divide communication sequences of stimuli and responses is referred to as punctuation (Watzlawick et al, 1967). This punctuation process implies that human communication should not only be represented by signals but also with communicative acts that represents the intuitive segmentation of human communication. Communicative acts can range from a spoken word to a segmented gesture (e.g., start and end time of a pointing) or a prosodic act (e.g., region of low pitch).

To improve the expressiveness of these communicative acts, the idea of encoding dictionary was proposed. Since communicative acts are not always synchronous, they are allowed to be represented with various delay and length. In the experiments with backchannel feedback, 13 encoding templates were identified to represent a wide range of ways that speaker actions can influence the listener backchannel feedback. These encoding templates will help to represent long-range dependencies that are otherwise hard to learn using directly a sequential probabilistic model (e.g., when the influence of an input feature decay slowly over time, possibly with a delay). An example of a long-range dependency will be the effect of low-pitch regions on backchannel feedback with an average delay of 0.7 seconds (observed by Ward and Tsukahara (2000)). In the prediction framework, the prediction model will pick an encoding template with a 0.5 seconds delay and the exact alignment will be learned by the sequential probabilistic model (e.g., Latent-Dynamic CRF) which will also take into account the influence of other input features. The three main types of encoding templates are:

- **Binary encoding:** This encoding is designed for speaker features which influence on listener backchannel is constraint to the duration of the speaker feature.
- **Step function:** This encoding is a generalization of binary encoding by adding two parameters: width of the encoded feature and delay between the start of the feature and its encoded version. This encoding is useful if the feature influence on backchannel is constant but with a certain delay and duration.
- **Ramp function:** This encoding linearly decreases for a set period of time (i.e., width parameter). This encoding is useful if the feature influence on backchannel is changing over time.

It is important to note that a feature can have an *individual* influence on backchannel and/or a *joint* influence. An *individual* influence means the

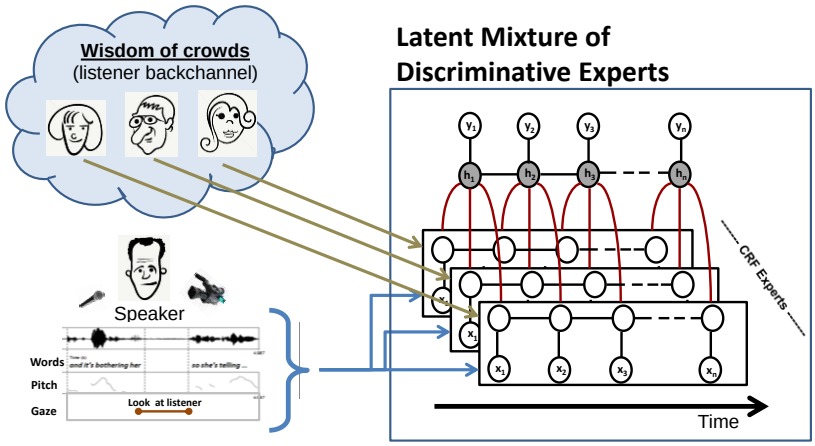

Figure 8: The approach for modeling wisdom of crowd: (1) multiple listeners experience the same series of stimuli (pre-recorded speakers) and (2) a Wisdom-LMDE model is learned using this wisdom of crowds, associating one expert for each listener.

input feature directly influences listener backchannel. For example, a long pause can by itself trigger backchannel feedback from the listener. A *joint* influence means that more than one feature is involved in triggering the feedback. For example, saying the word ``and'' followed by a look back at the listener can trigger listener feedback. This also means that a feature may need to be encoded more than one way since it may have an *individual* influence as well as one or more *joint* influences.

4.6 Wisdom of Crowds

In many real life scenarios, it is hard to collect the actual labels for training, because it is expensive or the labeling is subjective. To address this issue, a new direction of research appeared in the last decade, taking full advantage of the "wisdom of crowds" (Smith et al., 2005). In simple words, wisdom of crowds enables the fast acquisition of opinions from multiple annotators/experts.

Based on this intuition, wisdom of crowds was modeled using Parasocial Consensus Sampling paradigm (Huang et al., 2010) for data acquisition, which allows quided crowd members to experience the same situation. Parasocial Consensus Sampling (PCS) paradigm is based on the theory that people behave similarly when interacting through a media (e.g., video conference).

The goals of the computational model are to automatically discover the prototypical patterns of backchannel feedback and learn the dynamic between these patterns. This will allow the computational model to accurately predict the responses of a new listener even if he/she changes

Table 1: Comparison of the prediction model with previously published approaches. By integrating the knowledge from multiple listener, the Wisdom—LMDE is able to identify prototypical patterns in interpersonal dynamic.

Model	Wisdom of Crowds	Precision	Recall	F1-Score
Wisdom LMDE	Yes	0.2473	0.7349	**0.3701**
Consensus Classifier (Huang et al., 2010)	Yes	0.2217	0.3773	0.2793
CRF Mixture of Experts (Smith et al., 2005)	Yes	0.2696	0.4407	0.3345
AL Classifier(CRF)	No	0.2997	0.2819	0.2906
AL Classifier(LDCRF) (Morency et al., 2007)	No	0.1619	0.2996	0.2102
Multimodal LMDE (Ozkan et al., 2010)	No	0.2548	0.3752	0.3035
Random Classifier	No	0.1042	0.1250	0.1018
Rule Based Classifier(Ward et al.,2000)	No	0.1381	0.2195	0.1457

her backchannel patterns in the middle of the interaction. It will also improve generalization by allowing mixtures of these prototypical pattern.

To achieve these goals, a variant of the Latent Mixture of Discriminative Experts (Ozkan et al., 2010) was proposed to take full advantage of the wisdom of crowds. The Wisdom-LMDE model is based on a two step process: a Conditional Random Field (CRF) is learned first for each expert, and the outputs of these models are used as an input to an Latent Dynamic Conditional Random Field (LDCRF, see Figure 7) model, which is capable of learning the hidden structure within the input. In the Wisdom-LMDE, each expert corresponds to a different listener from the wisdom of crowds. Figure 8 show an overview of the approach.

Table 1 summarizes the experiments comparing the Wisdom-LMDE model with state-of-the-art approaches for behavior prediction. The Wisdom-LMDE model achieves the best f-1 score. The second best f-1 score is achieved by CRF Mixture of experts, which is the only model among other baseline models that combines the different listener labels in a late fusion manner. This result supports the claim that wisdom of clouds improves learning of prediction models.

5. Discussion

Modeling human communication dynamics enables the computational study of different aspect of human behaviors. While a backchannel feedback such as head nod may at first look like a conversational signal ("I acknowledge what you said"), it can also be interpreted as an emotional signal where the person is trying to show empathy or a social signal where the person is trying to show dominance by expressing a strong head nod. The complete study of human face-to-face communication needs to take into account these different types of signals: linguistic, conversational, emotional and social. In all four cases, the individual and interpersonal dynamics are

keys to a coherent interpretation.

As we already shown in this book chapter, modeling human communication dynamics is important for both recognition and prediction. One other important advantage of these computational models is the automatic analysis of human behaviors. Studying interactions is grueling and time-consuming work. The rule of thumb in the field is that each recorded minute of interaction takes an hour or more to analyze. Moreover, many social cues are subtle, and not easily noticed by even the most attentive psychologists.

By being able to automatically and efficiently analyze a large quantity of human interactions, and detect relevant patterns, these new tools will enable psychologists and linguists to find hidden behavioral patterns which may be too subtle for the human eye to detect, or may be just too rare during human interactions. A concrete example is the recent work which studied engagement and rapport between speakers and listeners, specifically examining a person's backchannel feedback during conversation (Ward & Tsukahara, 2000). This research revealed new predictive cues related to gaze shifts and specific spoken words which were not identified by previous psycho-linguistic studies. These results not only give an inspiration for future behavioral studies but also make possible a new generation of robots and virtual humans able to convey gestures and expressions at the appropriate times.

6. Bibliography

Feng, A.W., Huang, Y., Kallmann, M., Shapiro, A., (2012), An Analysis of Motion Blending Techniques, The Fifth International Conference on Motion in Games, 232-243

Bavelas, J. B., Coates, L., and Johnson, T. (2000), Listeners as Co-narrators, Journal of Personality and Social Psychology, 79(6):941-952

Burns, M. (1984). Rapport and relationships: The basis of child care. Journal of Child Care, 4:47–57.

D. DeVault, K. Sagae, and D. Traum. Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse, 2(1):143–170, 2011.

DeVito, J. (2008). The Interpersonal communication book. Pearson/Allyn and Bacon, 12th edition edition.

Drolet, A. and Morris, M. (2000), Rapport in conflict resolution: accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Experimental Social Psychology 36:26-50.

Fuchs D. (1987), Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education, 7:90-104.

Goldberg (2005), The secrets of successful mediators. Negotiation Journal, 21(3):365-376.

Hawley A.H (1950). Human ecology: A theory of community structure. Ronald Press.

Huang, L., Morency, L.-P., and Gratch, J. (2010). Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. AAMAS

Huang, Y., Kallmann, M. (2010) Motion Parameterization With Inverse Blending, Proceedings of the Third International Conference on Motion in Games

Kallmann, M., Marsella, S. (2005), Hierarchical Motion Controllers for Real-time Autonomous Virtual Humans, Proceedings of the 5th International Conference on Interactive Virtual Agents, 12-14

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. ThÃ³risson, and Hannes Vilhjalmsson, Towards a Common Framework for Multimodal Generation: The Behavior Markup Language, in *6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA, 2006.*

Kovar, L., and Gleicher, M. (2004), Automated Extraction and Parameterization of Motions in Large Data Sets, ACM Transactions on Graphics, Proceedings of SIGGRAPH 23(3):559-568

Lee, J. and Marsella, S.: Predicting speaker head nods and the effects of affective information. IEEE Transactions on Multimedia 12(6), 552 – 562 (October 2010).

Lee, J. and Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Proc. of the 6th Int. Conf. on Intelligent Virtual Agents. pp. 243–255 (2006)

Morency, L.-P., Quattoni, A., and Darrell, T. (2007), Latent-Dynamic Discriminative Models for Continuous Gesture Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June.

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2007), Head Gestures for Perceptual Interfaces: The Role of Context in Improving Recognition. Artificial Intelligence. Elsevier, 171(8-9):568-585, June.

Morency, L.-P., de Kok, I., and Gratch, J. (2008a), Predicting Listener Backchannels: A Probabilistic Multimodal Approach, Conference on Intelligent Virutal Agents

Ozkan, D., Sagae, K., and Morency, L.-P. (2010). Latent mixture of discriminative experts for multimodal prediction modeling. In International Conference on Computational Linguistics (COLING)

Pentland, A. (2004). Social dynamics: Signals and behavior. In IEEE Int. Conf. Developmental Learning, San Diego, CA, October.

Pettre, J. and Laumond, J.P. (2006), A Motion Capture-Based Control-Space Approach for Walking Mannequins, Computer Animation and Virtual Worlds, 17(2) 109-126

Rabiner, L.R. (1989): A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2):257-286

Rose, C., Bodenheimer, B., Cohen, M.F. (1998), Verbs and Adverbs: Multidimensional Motion Interpolation, IEEE Computer Graphics and Applications, 18 32-40

Shapiro, A. (2011) Building a Character Animation System, The Fourth International Conference on Motion in Games, 98-108

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C. (2004), Speaking With Hands: Creating Animated Conversational Characters from Recordings of Human Performance, ACM Transactions on Graphics, Proceedings of SIGGRAPH, 23(3) 506-513

Smith, A., Cohn, T., and Osborne, M. (2005). Logarithmic opinion pools for conditional random fields. In Annual Meeting of the Association for Computational Linguistics (ACL), pages 18–25

Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M. (2008), SmartBody: Behavior Realization for Embodied Conversational Agents, Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, 151-158

Tsui, P. and Schultz, G. (1985), Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients. American Journal of Orthopsychiatry 55:561-569

Ward, N. and Tsukahara, W. (2000), Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 23:1177-1207

Watzlawick, P., Bavelas, J. B., and Jackson, D. D. (1967). Pragmatics of Human Communication A Study of Interactional Patterns, Pathologies, and Paradoxes.