Appears in the International Conference on Interactive Virtual Agents, Kos Greece, 2005

Natural Behavior of a Listening Agent

R. M Maatman[1], Jonathan Gratch[2] and Stacy Marsella[3]

[1] University of Twente
Drienerlolaan 5, 7522 NB, Enschede, The Netherlands
[2] University of Southern California, Institute for Creative Technologies,
13274 Fiji Way, Marina del Rey, CA 90292, USA
[3] University of Southern California Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292, USA

Abstract. In contrast to the variety of listening behaviors produced in human-to-human interaction, most virtual agents sit or stand passively when a user speaks. This is a reflection of the fact that although the correct responsive behavior of a listener during a conversation is often related to the semantics, the state of current speech understanding technology is such that semantic information is unavailable until after an utterance is complete. This paper will illustrate that appropriate listening behavior can also be generated by other features of a speaker's behavior that are available in real time such as speech quality, posture shifts and head movements. This paper presents a mapping from these real-time obtainable features of a human speaker to agent listening behaviors.

1 Introduction

Have you ever presented in front of an unresponsive audience? Audiences that fail to react during a speech can negatively impact a speaker's performance, increasing cognitive load, raising doubts and breaking the speaker's rhythm. Not surprisingly, public speaking instructors often recommend that a speaker ignore unresponsive listeners. Listener behavior also has a critical impact in dyadic conversations (Warner 1996; Bernieri 1999; Lakin, Jefferis et al. 2003).

Similar effects have been demonstrated when people speak to virtual humans or avatars. As with human listeners, unresponsive virtual listeners can interfere with a speaker's cognitive processes. Worse, static characters can lead to a host of negative effects: Observers are more likely to criticize the quality of the graphics, they may feel less immersed, and they frequently form incorrect interpretations of the situation. For example, the authors of this paper have found in their own work that, in the context of highly emotional virtual scenarios, the lack of listener behavior can lead observers to read emotions into the virtual human's static behavior ("He must be really pissed.").

In the face of the important role that listener behavior plays, virtual human designers are faced with a basic dilemma. Most automated speech recognition (ASR) technologies used in virtual human systems do not give an interpretation of the speaker's utterance until the speaker is finished. There are no ongoing partial interpretations.

Therefore a listening virtual human cannot respond to what the speaker is saying as they speak. Rather, designers are forced to use simple behaviors at the start and end of an utterance (e.g., gaze or head nods), incorporate random (or "idle-time") listening behaviors, or more commonly have often ignored modeling listening behavior completely, instead focusing more on behavior of the agent while talking. Unfortunately, such design choices can be distracting or misread by the human participants.

The situation is not, however, as bleak as it may seem. The literature on human-human communication makes clear that there are solid correlations of listener behavior with various physical properties of the speaker's behavior, such as the speaker's nonverbal movement, the amplitude and pitch of the speech signal and key utterances. This suggests an approach to listening behavior that works in parallel with speech recognition. Namely, extracting information from the speech signal and physical movements of the human speaker that informs listener behavior as ASR is still processing the signal.

In this paper, we present such a system. After reviewing related work, we start by discussing the literature on listening behavior and how that literature informs our rules to drive virtual human listening behavior. The system we implemented is then detailed. Finally, we discuss our thoughts on evaluating the approach and present a preliminary evaluation.

2 Related ECA Research

The creation of human-appearing intelligent agents is an active area of computer science research. Known as *embodied conversational agents* (ECAs) or *virtual humans,* such systems allow humans to engage in face-to-face conversation with synthetic people, and attempt to model the full richness of such interactions including natural language communication, gestures, emotional expression, as well as the cognitive apparatus that underlies these capabilities (Cassell, Sullivan et al. 2000; Gratch, Rickel et al. 2002).

When it comes to conversational gestures, most virtual human research has focused on gestures related to speech production. Rea, for example, acts as a real estate agent (Cassell, Bickmore et al. 2000) and incorporates the Behavior Expression Animation Toolkit (BEAT) to automatically annotate virtual human speech with hand gestures, eye gaze, eyebrow movement, and intonation.

Some work has attempted to extract extra-linguistic features of a speakers' behavior, but not for the purpose of informing listening behaviors. For example, Brand's voice puppetry work attempts to learn a mapping between acoustic features and facial configurations to drive a virtual puppet with the speaker's voice. Several systems have attempted to recognize speaker gestures, though typically to help disambiguate speaker intent, as in "go that way [pointing left]". Such techniques could be repurposed to inform the present work.

Most virtual human systems have rudimentary listening behaviors triggered by the start and end of user speech. For example, the Mission Rehearsal Exercise system detects when a user begins speaking and orients its gaze towards the user for the duration of their speech, then looks away as it prepares to respond (Marsella, Gratch et al.

2003). These behaviors are typically fixed, however, and are not sensitive to the user's behavior during his or her utterance.

A few systems can condition their listening responses to features of the user's speech, though typically this feedback occurs only after an utterance is complete. For example, Neurobaby analyzes speech intonation and uses the extracted features to trigger emotional displays (Tosa 1993). More recently, Breazeal's Kismet system extracts emotional qualities in the user's speech (Breazeal and Aryananda 2002). Whenever the speech recognizer detects a pause in the speech, the previous utterance is classified (within one or two seconds) as indicating approval, an attentional bid, a prohibition, soothing or neutral. This is combined with Kismet's current emotional state to determine a facial expression and head posture. People who interact with Kismet often produce several utterances in succession, thus this approach is sufficient to provide a convincing illusion of real-time feedback.

Only a small number of systems have attempted to provide listening feedback *during* a user's utterance, and these methods have used only simple features of the speaker's behavior. For example, REA will execute a head nod or paraverbal (e.g. say "mm-hum") if the user pauses in mid-utterance for less than 500 milliseconds (Cassell, Bickmore et al. 1999). In contrast, a review of the psycholinguistic literature suggests that many other speaker behaviors are correlated with listener feedback and could be readily exploited by virtual characters.

3. Behavior of Human Listeners

The psycholinguistic literature has identified a variety of behaviors that listeners perform when in a conversation. Of course, many listener behaviors provide feedback about the semantic content the speaker's speech, but a large class of behaviors appear unrelated to specific meaning. Rather, these behaviors seem to trigger off of non-semantic features of the speaker's presentation, may precede complete understanding of the speech content, and are often generated without the listener or speaker's conscious awareness. Nonetheless, such behaviors can significantly influence the flow of a conversation and the impressions and feelings of the participants.

Here we review some of these behaviors, the circumstances that trigger their production and their hypothesized influence on the interaction. From this literature we extract a small number of simple rules that a listening agent could possibly utilize to drive its behavior.

3.1 Backchannel Continuers

Listeners frequently nod and utter paraverbals such as "uh-huh" and "mm-hmm" as someone is speaking. Within the psycholinguistic literature, such behaviors are referred to as *backchannel continuers* and are considered as a signal to the speaker that the communication is working and that they should continue speaking (Yngve 1970). Several researchers have developed models to predict when such feedback occurs. Cathcart, Carletta et al. (2003) propose a model based on pause duration and trigram part-of-speech frequency. According to the model of Ward and Tsukahara (2000),

backchannel continuers are associated with a lowering of pitch over some interval. Cassell (2000) argues that head nod's could result from the raised voice of the speaker. The approaches of Ward and Cassell are more amenable to a real-time treatment as they are based purely on simple properties of the audio signal, so we will adopt these methods for developing behavior mapping rules:

Rule-1: Lowering of pitch in speech signal → head nod

Rule-2: Raised loudness in speech signal → head nod

3.2 Disfluency

Spoken language often contains repetition, spurious words, pauses and filled pauses (e.g., ehm, um, un). Such disfluency is viewed as a signal to the listener that the speaker is experiencing processing problems or experiencing high cognitive load (Clark and Wasow 1998) and frequently elicit "take your time" feedback from the listener (Ward and Tsukahara 2000). According to own video analysis, rather than nodding or uttering sounds as in backchannel continuers, listeners tended to perform posture shifts, gaze shifts or frowns in response to disfluency. The presumed meaning of such a posture shift is the listener is telling the speaker to take his time (Cassell 2000). It should be possible to detect disfluency in the audio signal and this leads to the following behavior mapping rule:

Rule-3: Disfluency in speech signal → Posture Shift / Gaze shift / Frown

3.3 Mimicry

Listeners often mimic behavior of a speaker during a conversation. Although they are not necessarily aware of doing it, people in a conversation will adjust the rhythm of speech, their body posture and even their breathing to each other (Warner 1996; McFarland 2001; Lakin, Jefferis et al. 2003). Mimicry, when not exaggerated to the point of mocking, has a variety of positive influences on the interactants. Speakers who are mimicked are more helpful and generous toward the listener (Van baaren, Holland et al. 2004). Mimicry can result in the perception of a pleasant, natural conversation (Warner, Malloy et al. 1987). It may also be important in synchronizing conversational flow, for example, by providing expectations on when a speaker can be interrupted. Given such influences, many of the agent's listening behaviors should mimic aspects of the speaker's behavior.

One salient speaker behavior is shifts in posture. When a speaker shifts her posture, for example by changing her weight distribution from one leg to another, or by folding her arms, this is often mirrored by the listener. Such posture shifts, both for speakers and listeners, tend to occur at discourse segment boundaries and may function to help manage such transitions (Cassell, Nakano et al. 2001). When present, such mimicry has been shown to positively influence the emotional state of the

speaker (Van baaren, Holland et al. 2004). This suggests that a listening agent should detect posture shifts and mimic the resulting posture:

Rule-4: Speaker shifts posture → Mimic Posture

Gaze is also an important aspect of a speaker's behavior. Speakers will often gaze away from the listener, for example, when mentioning a concrete object within his vicinity, he will often look at it. When this lasts for a certain amount time, the listener could mimic this by looking in the same direction.

Rule-5: Speaker gazes away for longer period → Mimic Gaze

Listeners will frequently mimic the head gestures of a speaker. If a speaker shakes or nods his head, listeners may repeat this gesture. Although this may simply reflect an understanding and agreement with the speaker's utterance, many of us have probably been in conversations where such gestures were produced without any real understanding. In any event, an agent can easily mimic such gestures without explicit understanding. This leads to the following mimicry rule:

Rule-6: Speaker nods or shakes head → Mimic Head Gesture

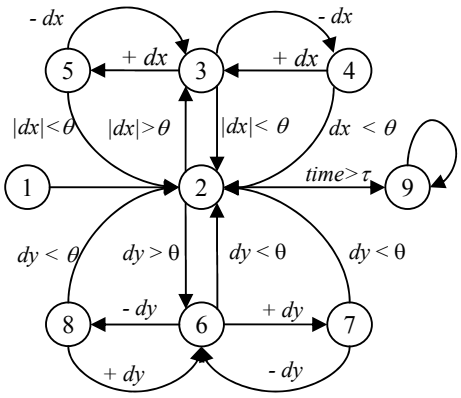
3.4 Other External Influences

Obviously, such rules are an oversimplification of the factors that mediate human behavior. Many factors influence the occurrence of gestures during a conversation. For example, listeners frequently mimic the facial expression of speakers and this apparently plays an important role in the perception of empathy (Sonnby-Borgstrom, Jonsson et al. 2003). Individuals also differ in their response to the same speech based on a variety of dispositional and situational factors. There are people who almost do not gesture at all and there are people who gesture like it is a workout. Often, this is related to the speaker's emotions during the conversation. For example, people tend to gesture more when excited and less when sad. Also, the relation of the two people is of importance. People tend to gesture remarkably more when they talk to a friend then when they are talking to a complete stranger (Welji and Duncan 2004).

Thus, the mapping presented here is not a complete coverage of all gestures that at all times are accompanied by the certain speech features, but could be sufficient to increase the perceived authenticity of the conversation.

4. Real-time Classification of Speaker Behavior

We implement the behavior rules listed above by detecting the various implicated aspects of the speaker's behavior. As such listening behaviors occur within utterances, this imposes strong real-time requirements on what features can be reasonably extracted given the limits of current technology. Here we describe the implementation of feature detectors that support the behavioral rules listed above.



1: Calibration state.
2: Head stationary.
3: Horizontal rotation. If follows state 4 or 5, report a shake.
4: Head looks right
5: Head looks left
6: Vertical rotation. If follows state 7 or 8, report a nod.
7: Head looks up
8: Head looks down
9: No move for τ seconds. Report a gaze.

Figure 1: State machine for detecting head gestures. *dx* (*dy)* denotes horizontal (vertical) rotation, θ is a distance and τ a time threshold..

There are two ways to determine physical features from a human in real time: image analysis or using 3d-trackers. Image analysis consists of recording the human with a camera and analyzing the data from the camera. The advantage of this method is that the complete human is visible and thus in theory all information could be extracted (with two or more cameras it might even be possible to get a 3D-image from the human). The disadvantages however are that it is computationally intensive and it is much work to create such a system from scratch.

In contrast to the image analysis method, there are tracker devices that can detect position and orientation with a high degree of accuracy. The advantage of using trackers is that they are fast and are not as computationally intensive as image analysis. The major drawback is that the trackers need to be set up and are only operational in a limited area. In addition to this, when using a tracker, only the point of the tracker is known and no other parts of the human body.

For this research however, a space was available where a tracker device was already operational and thus this device has been used to extract the physical features of the human. With this tracker it was possible to extract both head gestures (such as gazes, nodding and shaking) and posture shifts.

From the speech signal, we can extract certain features that are not directly related to the semantics of the speech. These features could then be calculated instantly from the input from a microphone. Only the basic features are considered here, because although the computational speed of the current computers is rapidly increasing, it should be kept relatively simple.

According to Milewski (1996), it is possible to extract frequency and intensity information from a speech signal in real time using a Fourier transformation. When these two aspects of the speech signal are known, many useful derivatives can be calculated, such as silences, monotone sounds, et cetera. Thus, when using this transformation, there can be much information available in real time concerning the features of the speech signal.

4.1 Detecting and Classifying Body Movements

4.1.1 Head Gestures

Certain head gestures and body posture are readily detected in real time through the use of a six-degree-of-freedom tracking sensor attached to the speakers head.

The speaker's head shakes, nods and gazes can be detected by attending to the orientation of the head over time. When the orientation of the head rotates back and forth along some axis, this indicates either a nod or a shake. It would be a shake it the movement is a horizontal rotation would be a nod if the movement is a vertical rotation. In contrast, if the head rotates to some orientation and holds this position for some time, it must mean that the speaker is staring in a certain direction. We implemented a finite state machine to recognize such speaker gestures (Figure 1).

This state diagram allows the system to extract the different gestures from the human head and thus let the agent perform some mimicking according to the mapping rules that have been specified in the gesture theory chapter. The relevant rules for the head gestures are numbers 5 and 6.

4.1.2 Posture Shifts

Certain posture shifts are obtained by using the tracker information. For example, if a speaker shifts her weight from one foot to the other, this is typically accompanied by the translation of the head with respect to a static position between the feet (see Figure Y). This way, a weight shift could be detected with just one tracker placed on top of the head of the speaker assuming they do not move their feet.

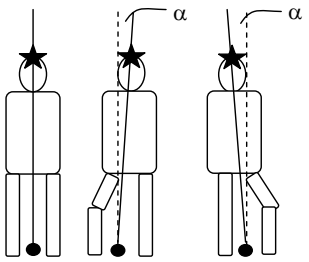


Figure 2: We detect posture shifts involving a weight shift by measuring angle α.

In our current system, we detect certain posture shifts in this way by measuring the angle α between the origin (dotted line in Figure Y) and the position of the tracker, placed on the head of the human. If this angle is greater than a certain threshold, this must mean the human is slouching. We use a restraining device to ensure that the speaker's feet remain stationary (this restriction could be eliminated be incorporating an additional tracker at the speaker's waist).

More specifically, posture shifts can be detected with just the angle between the head and the position between the legs and the height of the tracker (the length of the human), both obtainable from the tracker. With this, the angle α can be computed as follows:

$$\alpha = \text{atan}(dx / \text{height\_of\_tracker})$$

Where dx is the relative position of the tracker with respect to the initial position where the human is standing straight. The current angle is compared to the threshold in order to get the type of slouch (left, right or neutral).

4.2 Detecting and Classifying Acoustic Features

Besides the different gestures from the trackers, we also extract features from the speaker's audio signal such as pitch and loudness. Given the real-time requirements of the system, all audio feature detectors utilize the Fast Fourier Transform (FFT) which is not described here, but can be found in various sources of literature (e.g. Milewski, 1996). FFT can separate an audio signal into an arbitrary number of frequency-intensity pairs and these pairs can be used in the computation of several features of the speech signal. This means, that when the continuous speech signal is sampled in parts of a certain length, from each sample the FFT could be computed and compared.

4.2.1 Intensity Detection

With some minor adjustments, the algorithm by Arons (1994) is used to perform the intensity detection. After determining the average (normal) intensity of a speaker during an initialization phase, the real time intensity can be compared to a pre-computed threshold. This threshold would be the top one percent value of the initial intensity value computed during initialization. When the real time intensity value exceeds the threshold it must mean that there is a raise in intensity. With this information, mapping rule number 2 can be implemented.

Besides this approach to the computation of the intensity of a speech signal, another approach is proposed by Fernandez (2004). His model uses the computation of certain loudness features, which could improve the previously described method. The code to perform these loudness computations in Matlab has been available to us, but unfortunately this proved too slow to be useful for real time computations. Due to the limited amount of time available for this research, no optimizations could be made and thus the loudness detection model was discarded.

4.2.2 Pitch Detection

Pitch detection is done in a similar fashion as the intensity detection. For each sample, from the resulting frequency/intensity pairs of the FFT, the frequency is chosen that has the highest intensity and this is compared to the previous highest frequencies. This way, a significant drop or raise in the frequency of the speech can be detected.

We re-implemented the backchannel-algorithm by (Ward and Tsukahara 2000) within Matlab, using a pitch detection algorithm is available from the Matlab User Community. To actually implement this algorithm, the pitch values of the last 120 milliseconds have to be stored. Then, if all these values are below the 23rd percentile pitch level, an output can be performed after 700 milliseconds when all the other conditions are met. With this algorithm, mapping rule number 1 can be implemented.

4.2.3 Disfluency Detection

The final mapping rule depends on the detection of stuttering of disfluency in the speech signal. An example of this would be the expression 'uhhhh' for a longer period of time. To detect this, the frequency of the signal could be used. If a certain frequency holds for a longer period of time and does not vary much, this could mean disfluency.

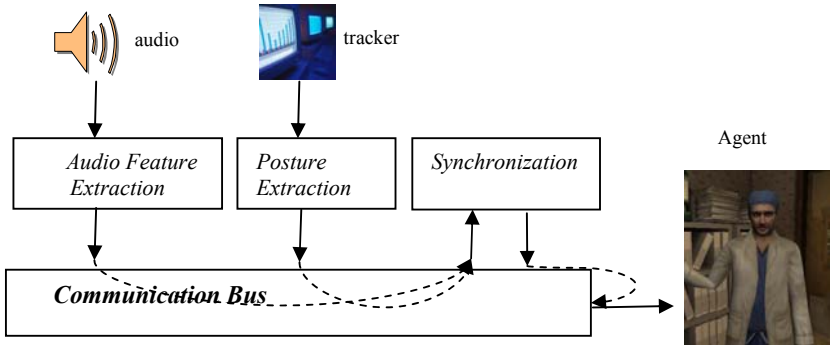


Figure 3. Overall system architecture

This method of detecting disfluency was proposed by Shriberg (1999), and concerns different types of disfluency, like filled/unfilled pauses, false starts and repetition of words. For this research, only the filled and unfilled pauses shall be considered, because these types do not depend on semantic information and thus can be detected using frequency and intensity respectively.

Specifically, Shriberg argues that a filled pause in an audio signal is accompanied by a relatively low frequency region for a period of at least 200 milliseconds. Thus, to extract this from the audio signal, the frequencies over 200 milliseconds have to be stored and evaluated. When the standard deviation of these frequencies is smaller than approximately one hertz, the module reports the detection of a disfluency.

5. Behavior of a Listening Agent

Through recognizing features of the speaker's behavior and applying these features to the behavior rules in Section 3, we can generate listening behaviors. The final issue is how to integrate these behaviors into an overall performance. As we are simultaneously recognizing features from multiple channels (head, body, ands speech), and listening behaviors have some duration, it is possible that a listening behavior will be proposed that overlaps with a currently executing behavior. This could easily result in unnatural behavior.

We use a synchronization module to solve this problem. This module parses messages on the communication bus and determines if the message is intended for the agent and which type of gesture is contained in the command. Once this parsing has been done, the function accompanying that type of gesture can be called. This function determines whether a previous gesture is still performing, and when this is not the case, a message is created which is to be sent to the agent. The module also incorporates a flexible behavior mapping allowing designers to easily experiment with alternative mappings between classes of listening behaviors and their physical manifestation.

6. Evaluation

This listening module could be integrated into a variety of embodied conversational agent applications, potentially improving the naturalness and subjective impressions of the interaction. In assessing the suitability of such an integration, we must consider several factors. Does the system correctly detect features of the speaker's behavior? Do the behavior mapping rules suggest appropriate feedback? Is the performed behavior judged natural at the time it is performed? Finally, do agent listening behaviors have the predicted influence on the human speaker's perceptions? Here we discuss the results of informal evaluations. Formal evaluations are planned for later this year.

In evaluating the system we adapt the "McNeill lab" paradigm (McNeill 1992) for studying gesture research. In this research, one participant, the Speaker, has previously observed some incident (e.g., a Sylvester and Tweety cartoon clip), and describes it to another participant, the Listener. Here, we replace the Listener with our agent system. Speakers see a life-sized animated character projected in a room. They stand in a foot restraining device in the middle of the room, wear a Intersense acoustic motion tracking sensor on their head and speak into a headset microphone. In the formal evaluation, we will use a 2x2 design. Speakers will be assigned to one of two priming conditions: they will be told that the Listener's behavior is either controlled by a human in another room or by a computer. The agent will either use our mapping rules or random behavior. Currently, we have performed preliminary evaluations using several staff members associated with the project. Notable findings from this initial feedback are reported here and will be used to adjust system parameters prior to the formal evaluations.

Backchannel continuers suggested by Ward's Pitch Detection Algorithm seem to occur in the appropriate location and thus Ward's algorithm does work fairly well. The only drawback here is that this algorithm is dependant of the initial recorded pitch threshold and thus when this initial recording would be modified, the results would be different.

The detection of disfluency consists of two parts, which are silences and filled pauses. The detection of silences worked very well although they tended to occur too soon. This can be resolved by extending the buffer which is an easy adaptation. The detection of filled pauses however did not work as well as predicted. The allowed variation in frequency of 0.05 Hertz proved too small and this has to be increased.

Observers reported the listening agent appeared more autonomous and natural. In particular, the occurrence of head nods and gazes seemed to contribute to this effect. Naturalness and autonomy does not necessarily translate into a feeling of engagement and our initial tests identified several factors that appeared to detract from engagement. For example, if the agent gazes away from the speaker too frequently, one is left with the impression that the agent is uninterested in the conversation. This led us to adjust downward the frequency of gaze shifts. We also decided the delay of 700 milliseconds in the algorithm by Ward was too long. This has been changed to 200 milliseconds and this led to more reasonable head nods. Some speakers systematically vary the intensity of the speech across utterances which can confuse our feature detectors. This could be resolved by re-computing the thresholds when the voice of the speaker undergoes a big change.

Although definitive testing has not been completed to date, the results of the informal tests seem promising. Especially the performing of head nods and the behavior when the human is silent seem to result in more natural behavior. Even though the speech signal was quite noisy, the responsive behavior improved the natural behavior of the agent.

7. Conclusions

Although they have not been a strong focus in the virtual human community, listening behaviors play an important role in promoting effective communication. A challenge in generating real-time listening behaviors is that the semantic content of a user's speech is typically available only after they are done speaking. Indeed, this information is sometimes only available a second or two after an utterance. This paper has reviewed the psycholinguistic literature to show that many listening behaviors are also correlated with physical behaviors that are easier to detect in real time. In other words, not only the meaning of the words is of importance, but also features as the intonation and the loudness of the speech signal.

Using this knowledge, we tried to find a mapping between these certain features and the accompanying gestures. The suggested mapping in this chapter is however not complete because there are many external factors that influence the occurrence of gestures, like the emotional state and relation of the persons involved. The suggested mapping can however be used to perform these gestures for example with a virtual human in order to let the virtual human react in a more natural way.

Acknowledgements

Stacy Marsella suggested the initial idea underlying this work – that some information in the speech signal could drive back channeling behaviors of a virtual character. Sue Duncan pointed us to several references on this literature and provided feedback on drafts. Roz Picard, Shri Narayanan and David Traum suggested several approaches for extracting information from the speech signal and Raul Galt provided us Matlab code that helped inform our work. Regina Cabrera provided valuable editorial feedback. This work was sponsored by the U. S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Arons, B. (1994). Pitch-Based Emphasis Detection For Segmenting Speech Recordings. International Conference on Spoken Language Processing.

Bernieri, J. E. G. a. F. J. (1999). "The Importance of Nonverbal Cues in Judging Rapport." Journal of Nonverbal Behavior **23**(4): 253-269.

Breazeal, C. and L. Aryananda (2002). "Recognition of Affective Communicative Intent in Robot-Directed Speech." Autonomous Robots **12**: 83-104.

Cassell, J. (2000). Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press**:** 1-27.

Cassell, J., T. Bickmore, et al. (1999). Embodiment in Conversational Interfaces: Rea. Conference on Human Factors in Computing Systems, Pittsburgh, PA.

Cassell, J., T. Bickmore, et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Boston, MIT Press**:** 29-63.

Cassell, J., Y. I. Nakano, et al. (2001). Non-verbal cues for discourse structure. Association for Computational Linguistics Joint EACL - ACL Conference.

Cassell, J., J. Sullivan, et al., Eds. (2000). Embodied Conversational Agents. Cambridge, MA, MIT Press.

Cathcart, N., J. Carletta, et al. (2003). A shallow model of backchannel continuers in spoken dialogue. 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest.

Clark, H. H. and T. Wasow (1998). "Repeating words in Spontaneous Speech." Cognitive Psychology**.37**: 204-242.

E., S. (1999). Phonetic Consequences of Speech Disfluency. International Congress of Phonetic Sciences, San Francisco, CA.

Fernandez, R. (2004). A Computational Model for the Automatic Recognition of Affect in Speech. Cambridge, MA, Ph.D. Thesis, MIT Media Arts and Science.

Gratch, J., J. Rickel, et al. (2002). Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems. **July/August:** 54-61.

Lakin, J. L., V. A. Jefferis, et al. (2003). "Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconsious Mimicry." Journal of Nonverbal Behavior **27**(3): 145-162.

Marsella, S., J. Gratch, et al. (2003). Expressive Behaviors for Virtual Worlds. Lifelike Characters Tools, Affective Functions and Applications. H. Prendinger and M. Ishizuka. Berlin, Springer-Verlag**:** 317-360.

McFarland, D. H. (2001). "Respiratory Markers of Conversational Interaction." Journal of Speech, Language, and Hearing Research **44**: 128-143.

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

Milewski, B. (1996). The Fourier Transform, Reliable Software, Relisoft.com.

Sonnby-Borgstrom, M., P. Jonsson, et al. (2003). "Emotional Empathy as Related to Mimicry Reactions at Different Levels of Information Processing." Journal of Nonverbal Behavior **27**(1): 3-23.

Tosa, N. (1993). "Neurobaby." ACM SIGGRAPH: 212-213.

Van baaren, R. B., R. W. Holland, et al. (2004). "Mimicry and Prosocial Behavior." Psychological Science **15**(1): 71-74.

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

Warner, R. (1996). Coordinated cycles in behavior and physiology during face-to-face social interactions. Dynamic patterns in communication processes. J. H. Watt and C. A. VanLear. Thousand Oaks, CA, SAGE publications.

Warner, R. M., D. Malloy, et al. (1987). "Rhythmic organization of social interaction and observer ratings of positive affect and involvement." Journal of Nonverbal Behavior **11**(2): 57-74.

Welji, H. and S. Duncan (2004). Characteristics of face-to-face interactions, with and without rapport: Friends vs. strangers. Symposium on Cognitive Processing Effects of 'Social Resonance' in Interaction, 26th Annual Meeting of the Cognitive Science Society.

Yngve, V. H. (1970). On getting a word in edgewise. Sixth regional Meeting of the Chicago Linguistic Society.