Proceedings of the 2000 IEEE International Workshop
on Robot and Human Interactive Communication
Osaka, Japan - September 27-29 2000

Proto-Conversations with an Anthropomorphic Robot

C. Breazeal
Department of Electrical Engineering and Computer Science
MIT Artificial Intelligence Lab
Cambridge, MA, 01239

Abstract

We present proto-dialog skills of an anthropomorphic robot, Kismet. The implementation is modeled after the proto-conversational skills of human infants. We have supplemented Kismet's vocal turn-taking skills with other para-linguistic social cues. We have found that naive subjects intuitively read and use these cues to regulate the exchange, making it smoother over time.

1 Introduction

Sociable machines are natural and intuitive for people to interact with and to teach. For the past three years, we have been exploring research issues in building socially intelligent humanoid robots [1]. This paper focuses on the challenge of building a robot that can engage in flexible and robust multi-modal turn-taking with people. A variety of animated conversation agents and a few conversational robots are under development at different labs [2], [3]. Many of these systems focus on adult-level discourse. In contrast, our work focuses on interactions that begin at an affective and physical level, and later develop to a linguistic level. Hence, similar to the proto-conversations that transpire between human infants and their caregivers [4], there is no linguistic content being communicated as of yet. However, the dynamics of the turn-taking exchange, and the use of gaze direction, facial display, and postural shifts closely resemble an animated dialog between robot and human.

We present the implementation of these proto-dialog skills on our anthropomorphic robot, Kismet. We focus on the implementation of the vocalization system and the behavior system. The robot can generate novel utterances with rich prosody. The utterances are accompanied with real-time lip sychronization and facial animation for emphasis. The proto-conversation skills are modeled after those of human infants (approximately 3 months of age). The turn-taking skills are supplemented with para-linguistic envelope displays that are used to regulate the rate of the exchange among people [5]. We present experimental results that illustrate the dynamics of the exchange. We also present data that suggests that naive subjects intuitively read the robot's para-linguistic social cues to entrain to the robot. As the result, the interaction becomes smoother over time as the robot and human tune to each other.

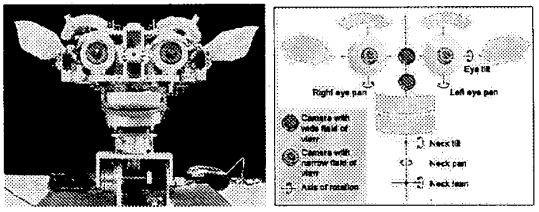

Figure 1: Kismet has a large set of expressive features – eyelids, eyebrows, ears, jaw, lips, neck and eye orientation. The schematic on the right shows the degrees of freedom relevant to visual perception (omitting the eyelids!). The eyes can turn independently along the horizontal (pan), but turn together along the vertical (tilt). The neck can turn the whole head horizontally and vertically, and can also crane forward. Two cameras with narrow "foveal" fields of view rotate with the eyes. Two central cameras with wide fields of view rotate with the neck. These cameras are unaffected by the orientation of the eyes.

2 Vocalization System

We have implemented an expressive vocalization system that supports novel utterances. In doing so, we have addressed issues regarding the expressiveness and richness of Kismet's vocal modality, and how it supports social interaction. We have found that the vocal utterances are rich enough to facilitate interesting proto-dialogs with people.

The robot's vocalization capabilities are generated through a commercial articulatory synthesizer, *DECtalk v4.5*. The parameters of the synthesizer model are based on the physiological characteristics of the human articulatory tract. Adjustment of the articulatory parameters make it possible to convey emotional states through vocalizations [6], and has been implemented on Kismet [1]. It is also possible to convey personality by designing a custom voice for the robot. As such, Kismet's voice is that of a young child. The software can accept strings of phonemes along with commands to specify the pitch and timing of the utterance. Hence, Kismet's vocalization system generates both phoneme strings and command settings, and says them in near real-time. The synthe-

0-7803-6273-X/00/$10.00©2000 IEEE

sizer also extracts phoneme and pitch information which are used to coordinate real-time lip synchronization.

2.1 Generating the Utterance

To engage in proto-dialogs with its human caregiver and to partake in vocal play, Kismet must be able to generate its own utterances. Based upon DECtalk's phonemic speech mode, the generated string to be synthesized is assembled from pitch accents, phonemes, and end syntax. The end syntax is a requirement of DECtalk and does not serve a grammatical function. However, as with the pitch accents, it does influence the prosody of the utterance and is used in this manner. The algorithm outlined below produces a style of speech that is reminiscent of a tonal dialect. As it stands, it is quite distinctive and contributes significantly to Kismet's personality (as it pertains to its manner of vocal babbling). We are currently working on having Kismet adjust its utterance based on what it hears.

Randomly choose number of proto-words,
$getUtteranceLength() = length_{utterance}$
For $i = (0, length_{utterance})$, generate a proto-word, $protoWord$
  Generate a $(wordAccent, word)$ pair
  Randomly choose word accent, $getAccent()$
  Randomly choose number of syllables of proto-word, $getWordLength() = length_{word}$
  Choose which syllable receives primary stress, $assignStress()$
  For $j = (0, length_{word})$ generate a syllable
    Randomly choose the type of syllable, $syllableType$
    if $syllableType = vowelOnly$
      if this syllable has primary stress
      then $syllable = getStress() + getVowel()+ getDuration()$
      else $syllable = getVowel() + getDuration()$
    if $syllableType = consonantVowel$
      if this syllable has primary stress
      then $syllable = getConsonant() + getStress()+ getVowel() + getDuration()$
      else $syllable = getConsonant() + getVowel()+ getDuration()$
    if $syllableType = consonantVowelConsonant$
      if this syllable has primary stress
      then $syllable = getConsonant() + getStress()+ getVowel() + getDuration()+ getConsonant()$
      else $syllable = getConsonant() + getVowel()+ getDuration() + getConsonant()$
    if $syllableType = vowelVowel$
      if this syllable has primary stress
      then $syllable = getStress() + getVowel()+ getDuration() + getvowel()+ getDuration()$
      else $syllable = getVowel()+ getDuration()+ getVowel() + getDuration()$
    $protoWord = append(protoWord, syllable)$
  $protoWord = append(wordAccent, protoWord)$
$utterance = append(utterance, protoWord)$

Where:

- $GetUtteranceLength()$ randomly chooses a number between $(1, 5)$. This specifies the number of proto-words in a given utterance.
- $GetWordLength()$ randomly chooses a number between $(1, 3)$. This specifies the number of syllables in a given proto-word.
- $GetPunctuation()$ randomly chooses one of end syntax markers. This is biased by emotional state to influence the end of the pitch contour.
- $GetAccent()$ randomly choose one of six accents (including no accent).
- $assignStress()$ selects which syllable receives primary stress.
- $getVowel()$ randomly choose one of eighteen vowel phonemes.
- $getConsonant()$ randomly chooses one of twenty-six consonant phonemes.
- $getStress()$ gets the primary stress accent.
- $getDuration()$ randomly chooses a number between $(100, 500)$ that specifies the vowel duration in msec. This selection is biased by the emotional state where lower arousal vowels tend to have longer duration, and high arousal states have shorter duration.

2.2 Real-Time Lip Synchronization and Facial Animation

Given Kismet's ability to express itself vocally, it is important that the robot also be able to support this vocal channel with coordinated facial animation. This includes synchronized lip movements to accompany speech along with facial animation to lend additional emphasis to the stressed syllables. These complementary motor modalities greatly enhance the robot's delivery when it speaks, giving the impression that the robot "means" what it says. This makes the interaction more engaging for the human and facilitates proto-dialog.

To implement lip synchronization on Kismet, a variety of information must be computed in real-time from the speech signal. By placing DECtalk in *memory mode* and issuing the command string (utterance with synthesizer settings), the DECtalk software generates the speech waveform and writes it to memory (a 11.025

kHz waveform). In addition, DECtalk extracts time-stamped phoneme information. From the speech waveform, we compute its time-varying energy over a window size of 335 samples. We take care to synchronize the phoneme and energy information, and send $(phoneme(t), energy(t))$ pairs to the PC computer at 33 Hz to coordinate jaw and lip motor control.

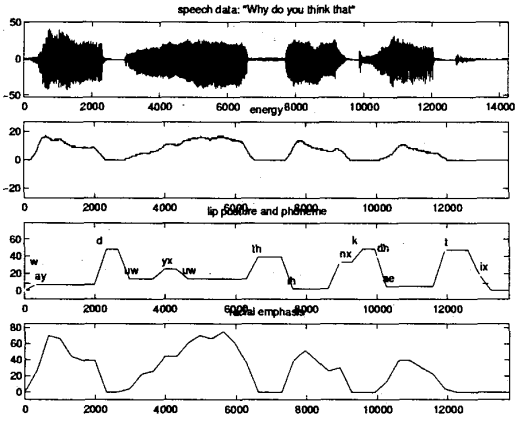

Figure 2: Plot of speech signal, energy, phonemes/lip posture, and facial emphasis for the phrase "Why do you think that". Time is in 0.1 ms increments. The total amount of time to vocalize the phrase is 1.4 sec.

To control the jaw, the PC receives the phoneme and energy information and updates the commanded jaw position at 10 Hz. The mapping from energy to jaw opening is linear, bounded within a range where the minimum position corresponds to a closed mouth, and the maximum position corresponds to an open mouth characteristic of surprise. Using only energy to control jaw position produces a lively effect but has its limitations [7]. For Kismet, the phoneme information is used to make sure that the jaw is closed when either a m, p, or b is spoken or there is silence. This may not necessarily be the case if only energy were used.

Upon receiving the phoneme and energy information from the vocalization system, the *vocal communication* process passes this information to the motor skill system which converts the energy information into a measure of facial emphasis (linearly scaling the energy), which is then passed onto the lip synchronization and facial animation processes of the face control motor system. The motor skill system also maps the phoneme information onto lip postures and passes this information to the *lip synchronization* and *facial animation* processes of the motor system that controls the face.

Lip synchronization is only part of the equation, however. Faces are not completely still when speaking, but move in synchrony to provide emphasis along with the speech. Using the energy of the speech signal to animate Kismet's face (along with the lips and jaw) greatly enhances the impression that Kismet "means" what it says. For Kismet, the energy ofthe speech signal influences the movement of its eyelids and ears. Larger speech amplitudes result in a proportional widening of the eyes and downward pulse of the ears. This adds a nice degree of facial emphasis to accompany the stress of the vocalization.

3 Proto-Conversation Behaviors

With respect to social interaction, Kismet's behavior system supports proto-dialogs with a humans, reminiscent of three-month-old infants with their caregiver. Tronick and his collaborators have identified five phases that characterize such exchanges: *initiation, mutual-orientation, greeting, play-dialog,* and *disengagement* [4]. Each phase represents a collection of behaviors which mark the state of the communication, and every phase is present in every interaction. We have implemented these phases into Kismet's proto-conversation behaviors to capture a similar social dynamic with humans.

The related behaviors reside within the `social-play` behavior group (see the lower left cluster of figure 3). This behavior group encapsulates Kismet's engagement strategies for establishing proto-dialogs during face-to-face exchanges. The turn-taking behavior is supplemented with envelope displays. As discussed in Cassell (2000), these paralinguistic social cues (such as raising of the brows at the end of a turn, or averting gaze at the start of a turn) are used to humans to regulate the exchange of speaking turns. These cues are particularly important for Kismet because processing limitations force the robot to take-turns at a slower rate than is typical for human adults. However, humans seem to intuitively read Kismet's cues and use them to regulate the rate of exchange at a pace where both partners perform well.

3.1 Calling Behavior

The first engagement task is the `call-to-person` behavior. This behavior is relevant when a person is in view of the robot but too far for face-to-face exchange. The goal of the behavior is to lure the person into face-to-face interaction range (ideally, about three feet from the robot). To accomplish this, Kismet sends a social cue, the `calling` display, directed to the person within calling range.

The releaser affiliated with this behavior combines skin-tone with proximity measures. It fires when the person is four to seven feet from the robot. The human observer sees the robot orient towards him/her, crane its neck forward, wiggle its ears with large amplitude movements, and vocalize excitedly. The display is designed to attract a person's attention. The robot then resumes a neutral posture, perks its ears, and raises its brows in an expectant manner. It waits in this posture for a while,

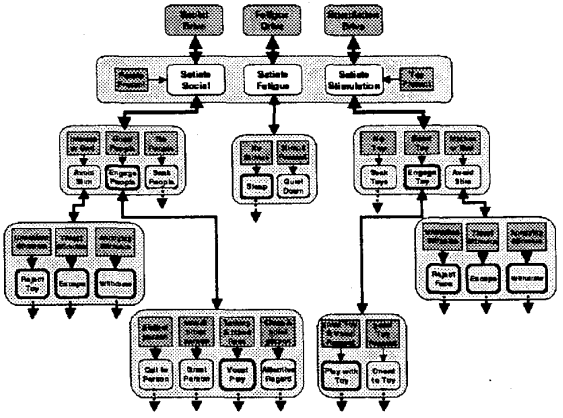

Figure 3: Kismet's behavior hierarchy. See text for a description.

giving the person time to approach before the calling sequence resumes. The `call-to-person` behavior will continue to request the display from the motor system until it is either successful and becomes deactivated, or it becomes irrelevant.

3.2 Greeting Behavior

The second task is the `greet-person` behavior. This behavior is relevant when the person has just entered into face-to-face interaction range. It is also relevant if the `social-play` behavior group has just become active and a person is already within face-to-face range. The goal of the behavior is to socially acknowledge the human and to initiate a close interaction. When active, it makes a request of the motor system to perform the greeting display. The display involves making eye contact with the person and smiling at them while waving the ears gently. It often immediately follows the success of the `call-to-person` behavior. It is a transient response, only issued once, as its completion signals the success of this behavior.

3.3 Attentive-Regard Behavior

The third task is `attentive-regard`. This behavior is active when the person has already established a good face-to-face interaction distance with the robot but remains silent. The goal of the behavior is to visually attend to the person and to appear open to interaction. To accomplish this, it sends a request to the motor system to hold gaze on the person, ideally looking into the person's eyes if the eye detector can locate them. The robot watches the person intently and vocalizes occasionally. If the person does speak, this behavior loses the competition to the `vocal-play` behavior.

3.4 Turn-taking Behavior

The forth task is `vocal-play`. The goal of this behavior is to carry out a proto-dialog with the person. It is relevant when the person is within face-to-face interaction distance and has spoken. To perform this task successfully, the `vocal-play` behavior must closely regulate turn-taking with the human. This involves a close interaction with the perceptual system to perceive the relevant turn-taking cues from the person (i.e., that a person is present and whether or not there is speech occurring), and with the motor system to send the relevant turn-taking cues back to the person.

There are four turn-taking phases this behavior must recognize and respond to. Each state is recognized using distinct perceptual cues, and each phase involves making specific display requests of the motor system.

- *Relinquish speaking turn*: This phase is entered immediately after the robot finishes speaking. The robot relinquishes its turn by craning its neck forward, raising its brows, and making eye-contact (in adult humans, shifting gaze direction is sufficient, but we exaggerated the display for Kismet to increase its readability). It holds its gaze on the person throughout this phase. However, due to noise in the visual system, in practice the eyes tend to flit about the person's face, perhaps even leaving it briefly and then returning soon afterwards. This display signals that the robot has finished speaking and is waiting for the human to say something. It will time out after a few seconds (approx. 8 seconds) if the person does not respond. At this point, the robot reacquires its turn and issues another vocalization in an attempt to reinitiate the dialog.

- *Attend to human's speech*: Once the perceptual system acknowledges that the human has started speaking, the robot's ears perk. This little feedback cue signals that the robot is listening to the person speak. The robot looks generally attentive to the person and continues to maintain eye contact if possible.

- *Reacquire speaking turn*: This phase is entered when the perceptual system acknowledges that the person's speech has ended. The robot signals that it is about to speak by leaning back to a neutral posture and averting its gaze. The robot is likely to blink its eyes as it shifts posture.

- *Deliver speech*: Soon after the robot shifts its posture back to neutral, the robot vocalizes. The utterances are short babbles, generated by the vocalization system (presented in section 2). Sometimes more than one is issued. The eyes migrate back to the person's face, to their eyes if possible. Just before the robot is prepared to finish this phase, it is likely to blink. The behavior transitions back to the relinquish turn phase and the cycle resumes.

The system is designed to maintain social exchanges with a person for about twenty minutes; at this point the other drives typically begin to dominate the robot's motivation. When this occurs, the robot begins to behave in a fussy manner – the robot becomes more distracted by other things around it, and it makes fussy faces more frequently. It is more difficult to engage in proto-dialog. Overall, it is a significant change in behavior. People seem to readily sense the change and try to vary the interaction, often by introducing a toy. The smile that appears on the robot's face, and the level of attention that it pays to the toy, are strong cues that the robot is now involved in satiating its **stimulation drive**.

4 Experiments and Analysis

The behavior system produces interaction dynamics that are similar to the five phases of infant social interactions (initiation, mutual-orientation, greeting, play-dialog, and disengagement) discussed in section 3. As presented in [8], these dynamic phases are not explicitly represented in the behavior system, but emerge from the interaction of the synthetic nervous system with the environment. By producing behaviors akin to the proto-social responses of human infants, we exploit the caregivers natural tendencies to treat the robot as a social creature, and thus to respond in characteristic ways to the robot's overtures. This reliance on the external world produces dynamic behavior that is both flexible and robust.

Figure 4 shows Kismet's dynamic responses during face-to-face interaction with a caregiver. Kismet is initially looking for a person and displaying sadness (the initiation phase). The sad expression evokes nurturing responses from the caregiver. The robot begins moving its eyes looking for a face stimulus ($t < 8$). When it finds the caregiver's face, it makes a large eye movement to enter into mutual regard ($t \approx 10$). Once the face is foveated, the robot displays a greeting behavior by wiggling its ears ($t \approx 11$), and begins a play-dialog phase of interaction with the caregiver ($t > 12$). Kismet continues to engage the caregiver until the caregiver moves outside the field of view ($t \approx 28$). Kismet quickly becomes sad, and begins to search for a face, which it re-acquires when the caregiver returns ($t \approx 42$). Eventually, the robot habituates to the interaction with the caregiver and begins to attend to a toy that the caregiver has provided ($60 < t < 75$). While interacting with the toy, the robot displays interest and moves its eyes to follow the moving toy. Kismet soon habituates to this stimulus, and returns to its play-dialog with the caregiver ($75 < t < 100$). A final disengagement phase occurs ($t \approx 100$) when the robot's attention shifts back to the toy.

4.1 Turn-Taking Experiments

Within the play-dialog phase, Kismet employs different social cues to regulate the rate of vocal exchanges.

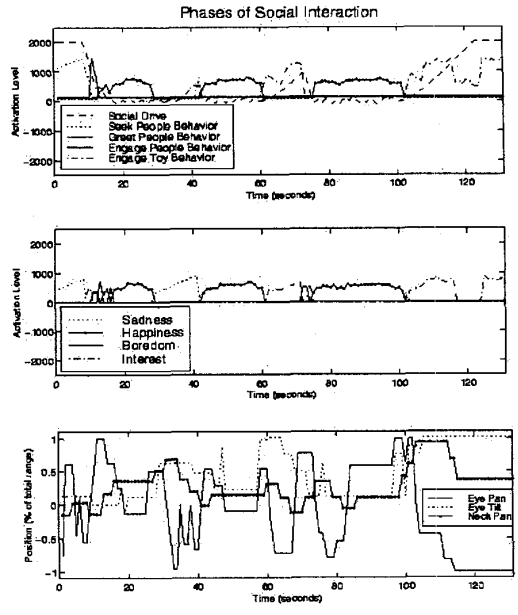

Figure 4: Cyclic responses during social interaction. Behaviors and drives (top), emotions (middle), and motor output (bottom) are plotted for a single trial of approximately 130 seconds. See text for description.

These include both eye movements as well as postural and facial displays. These cues encourage the subjects to slow down and shorten their speech. This benefits the auditory processing capabilities of the robot.

To investigate Kismet's performance in engaging people in proto-dialogs, we invited three naive subjects to interact with Kismet. They ranged in age from 25 to 28 years of age. There was one male and two females, all professionals. They were simply asked to talk to the robot. Their interactions were video recorded for further analysis.

Often the subjects begin the session by speaking longer phrases and only using the robot's vocal behavior to gauge their speaking turn. They also expect the robot to respond immediately after they finish talking. Within the first couple of exchanges, they may notice that the robot interrupts them, and they begin to adapt to Kismet's rate. They start to use shorter phrases, wait longer for the robot to respond, and more carefully watch the robot's turn taking cues. The robot prompts the other for their turn by craning it's neck forward, raising it's brows, and looking at the person's face when its ready for them to speak. It will hold this posture for a few seconds until the person responds. Often, within a second of this display, the subject does so. The robot then leans back to a neutral posture, assumes a neutral expression, and tends to shift its gaze away from the per-

		time stamp (min:sec)	time between disturbances (sec)
subject 1	start @ 15:20	15:20 – 15:33	13
		15:37 – 15:54	21
		15:56 – 16:15	19
		16:20 – 17:25	70
	end @ 18:07	17:30 – 18:07	37+
subject 2	start @ 6:43	6:43 – 6:50	7
		6:54 – 7:15	21
		7:18 – 8:02	44
	end @ 8:43	8:06 – 8:43	37+
subject 3	start @ 4:52 min	4:52 – 4:58	10
		5:08 – 5:23	15
		5:30 – 5:54	24
		6:00 – 6:53	53
		6:58 – 7:16	18
		7:18 – 8:16	58
		8:25 – 9:10	45
	end @ 10:40 min	9:20 – 10:40	80+

Figure 5: Data illustrating evidence for entrainment of human to robot.

	subject 1		subject 2		subject 3		average
	data	percentage	data	percentage	data	percentage	
clean turns	35	83%	45	85%	83	78%	82%
interrupts	4	10%	4	7.5%	16	15%	11%
prompts	3	7%	4	7.5%	7	7%	7%
significant flow disturbances	3	7%	3	5.7%	7	7%	6.5%
total speaking turns	42		53		106		

Figure 6: Kismet's turn taking performance during proto-dialog with three naive subjects. Significant disturbances are small clusters of pauses and interruptions between Kismet and the subject until turn-taking become coordinated again.

son. This cue indicates that the robot is about to speak. The robot typically issues one utterance, but it may issue several. Nonetheless, as the exchange proceeds, the subjects tend to wait until prompted.

Before the subjects adapt their behavior to the robot's capabilities, the robot is more likely to interrupt them. There tend to be more frequent delays in the flow of "conversation" where the human prompts the robot again for a response. Often these "hiccups" in the flow appear in short clusters of mutual interruptions and pauses (often over 2 to 4 speaking turns) before the turns become coordinated and the flow smoothes out. However, by analyzing the video of these human-robot "conversations", there is evidence that people entrain to the robot (see figure 5). These "hiccups" become less frequent. The human and robot are able to carry on longer sequences of clean turn transitions. At this point the rate of vocal exchange is well matched to the robot's perceptual limitations. The vocal exchange is reasonably fluid. Table 6 shows that the robot is engaged in a smooth proto-dialog with the human partner the majority of the time (about 82%).

5 Summary

Kismet can engage a human in compelling social interaction during face-to-face exchanges. People seem to interpret Kismet's emotive responses quite naturally and adjust their behavior so that it is suitable for the robot. Furthermore, people seem to entrain to the robot by reading its turn-taking cues. As a result, the interactions become smoother over time.

Acknowledgments

This work was supported by ONR and DARPA under MURI N00014-95-1-0600, and by DARPA under contract DABT 63-99-1-0012. The author would like to acknowledge the contributions of Paul Fitzpatrick and Brian Scassellati to Kismet's visual system. Jim Glass and Lee Hetherington of the Spoken Language Systems Group at MIT generously ported the Sapphire speech recognition software to Kismet.

References

[1] C. Breazeal, *Sociable Machines: Expressive Social Exchange Between Humans and Robots*, Sc.D. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA. 2000.

[2] J. Cassell, J. Sullivan, S. Prevost & E. Churchill, *Embodied Conversation Agents*, MIT Press, Cambridge, MA. 2000.

[3] Y. Matsusaka & T. Kobayashi, "Human interface of humanoid robot realizing group communication in real space," *Proceedings if HURO99*, Tokyo, Japan. pp. 188–193, 1999.

[4] E. Tronick, H. Als & L. Adamson, "Structure of early face-to-face communication interactions", in M. Bullowa ed., *Before Speech*, Cambridge University Press, pp. 349–370, 1979.

[5] J. Cassell, "Nudge Nudge Wink Wink: Elements of face-to-face conversation for embodied conversational agents", in J. Cassell, J. Sullivan, S. Prevost & E. Churchill eds., *Embodied Conversational Agents*, MIT Press, Cambridge, MA. 2000.

[6] J. Cahn, *Generating Expression in Synthesized Speech*, Masters Thesis, MIT Media Lab, Cambridge, MA. 1990.

[7] F. Parke & K. Waters, *Computer Facial Animation*, A. K. Peters, Wellesley, MA. 1996.

[8] C. Breazeal & B. Scassellati, "How to build robots that make friends and influence people," *Proceedings if IROS99*, Kyonju, Korea. pp. 858–863, 1999.