25th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
August 26-31, 2016. Columbia University, NY, USA

ERICA: The ERATO Intelligent Conversational Android

Dylan F. Glas, Takashi Minato, Carlos T. Ishi,
Tatsuya Kawahara, *Senior Member, IEEE,* and Hiroshi Ishiguro, *Member, IEEE*

***Abstract*—The development of an android with convincingly lifelike appearance and behavior has been a long-standing goal in robotics, and recent years have seen great progress in many of the technologies needed to create such androids. However, it is necessary to actually integrate these technologies into a robot system in order to assess the progress that has been made towards this goal and to identify important areas for future work. To this end, we are developing ERICA, an autonomous android system capable of conversational interaction, featuring advanced sensing and speech synthesis technologies, and arguably the most humanlike android built to date. Although the project is ongoing, initial development of the basic android platform has been completed. In this paper we present an overview of the requirements and design of the platform, describe the development process of an interactive application, report on ERICA's first autonomous public demonstration, and discuss the main technical challenges that remain to be addressed in order to create humanlike, autonomous androids.**

I. Introduction

In recent years, androids have become increasingly visible in both research and the popular media. Android replicas of celebrities and individuals are appearing in the news, and androids are depicted in film and television living and working alongside people in daily life. However, many contemporary androids are fully or partially teleoperated [1], and generally speaking, today's androids are often very limited in their ability to conduct autonomous conversational interactions. Despite the excitement over the appearance of these robots, it is important not to lose sight of the goal shared by many, of creating fully-autonomous interactive androids.

A. System Integration

The challenge of developing a conversational android system is not merely a simple problem of plugging together off-the-shelf components. There are interdependencies between components, and elements of the architecture need to be customized and tuned based on the environment and system configuration. Most works in this field focus on individual technologies, such as speech or gesture recognition, or on psychological studies of appearance or behavior, and issues of architecture design and system integration are not well-represented in research. However, it is not only important to share these techniques and designs, but we believe that the process of actually building and integrating an autonomous android will enable us to better understand the progress we have made and identify the important challenges remaining to achieve the goal of creating a convincing artificial human.

Figure 1. Photograph of ERICA, the android platform presented in this work.

In this work, we will present the architecture of ERICA (Fig. 1), an autonomous conversational android with an unprecedented level of humanlikeness and expressivity, as well as advanced multimodal sensing capabilities. We will discuss the requirements and component technologies of the system, and we will present an example of how we created an interactive application for a public demonstration of the platform's capabilities. Since no consensus yet exists as to how the architecture for an autonomous android should be designed, it is our hope that sharing these designs and experiences will be of value to the android community.

B. Related Work

In comparison with other humanoid robots, androids face a set of unique challenges. For example, to avoid "uncanny valley" effects [2], the generation of extremely humanlike behavior is of great importance. In this section we will discuss other social robots and how they compare with ERICA.

1) Non-humanoid robots and virtual agents

Highly realistic virtual agents have been created that can interact conversationally. The Virtual Human Toolkit [3] provides a set of tools for dialog and character design for photorealistic animated graphical avatars. However, graphical agents are free from many real-world mechanical constraints, and thus lack some degree of physical realism. Furhat [4] is a robot that begins to bridge the gap between 2D and 3D, consisting of a movable head with a back-projected face, enabling a wide range of facial expressions.

*Research supported by the JST ERATO Ishiguro Symbiotic Human Robot Interaction Project.

D. F. Glas, T. Minato, C. T. Ishi, and H. Ishiguro are with the Hiroshi Ishiguro Laboratories at ATR, Kyoto, Japan, 619-0288, and the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

T. Kawahara is with School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan, 606-8501, and the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

(Corresponding author: D. F. Glas, phone: +81-774-95-1405; fax: +81-774-95-1408; e-mail: dylan@atr.jp)

2) Humanoid robots

Several humanoid robots with varying degrees of anthropomorphism have been developed which are humanlike enough to conduct interesting interactions using natural gestures and other social cues. These robots often take mechanical, animal, cartoonlike, or abstract forms.

Leonardo [5] is an example of a highly-expressive robot designed for human interaction research. Its 65 degrees of freedom enable highly expressive animations, and although its behaviors are frequently scripted, framework components have been developed for handling key functions like natural language understanding, spatial reasoning, and attention. However, its form is that of a small animal or plush toy.

Similarly, Aldebaran's Nao robot [1] is a widely-used platform for human-robot interaction research, and their recently released robot Pepper[2] is another promising platform for rich interactive human-robot communication. However, the design of both robots is mechanical in nature rather than biological. Robovie [6] and Simon [7] are other humanoid robot platforms with mechanical-style designs which have been used extensively in human-robot interaction research.

Giving a robot a humanoid rather than strictly humanlike appearance can be a useful and effective strategy for avoiding the uncanny valley effect. However, this very fact makes them inappropriate for use as platforms for studying the challenges of creating a realistic humanlike android.

3) Androids

A variety of lifelike androids have already been developed. Hanson Robotics has produced many highly-expressive human head robots, such as the PKD robot [8], BINA48, Han, and Jules, some of which have been placed on bodies. These robots exhibit advanced AI techniques and highly articulated facial expressions, but they often look robotic, sometimes with metallic parts or exposed wires, and generally lack expressive speech synthesis. The Geminoid series of androids [9] also feature highly humanlike appearance and expressivity [10].

ERICA's physical appearance was designed to improve on the humanlike appearance of recent androids like Otonaroid and Kodomoroid. Careful attention has also been given to the design of her expressive and humanlike speech synthesis. For these reasons, we feel confident that ERICA is the most humanlike android developed to date.

C. The ERICA Platform

The android platform we have created in this work is called "ERICA", an acronym for "ERATO Intelligent Conversational Android", named for the ERATO Ishiguro Symbiotic Human-Robot Interaction Project, a collaborative research effort between teams at ATR, Osaka University, and Kyoto University, with the goal of developing a fully-autonomous android. Three identical copies of the robot have been created, one at each of the participating institutions.

With a highly lifelike appearance, smooth motion control, expressive speech synthesis, state-of-the-art multimodal sensor systems, and a control architecture designed to support dialog management and nonverbal behaviors, ERICA is meant to showcase the most advanced elements of modern android technology. Unlike many existing androids, ERICA is not modeled after a real person or designed for telepresence applications. Instead, she is designed as a general research platform for android autonomy.

At the moment, the project is in an early phase, so ERICA's high-level conversational capabilities are still quite limited. However, the robot platform has been completed, and ERICA was recently unveiled in a public demonstration showcasing the basic functionalities of the platform. In the demonstration, she autonomously answered questions from reporters and conducted conversations with other presenters while demonstrating a rich set of nonverbal behaviors.

We will present the robot platform in Sec. II and the process of developing an application for ERICA in Sec. III. The public demonstration will be described in Sec. IV, and we will discuss our progress and areas for future work in Sec. V.

II. Platform Architecture

A. Hardware and Actuation

Lifelike appearance and motion are essential requirements for humanlike androids. The mechanical and aesthetic design of ERICA were developed together with the android manufacturer A-Lab[3]. Although ERICA's overall design is similar to other androids produced by A-Lab, there are a few important improvements over earlier models.

1) External appearance

The design policy used for creating ERICA's face was similar to that used for the Telenoid robot [11, 12]. The features were kept neutral to allow a wide range of potential applications. In order to create an attractive robot that people would feel comfortable interacting with, the face was designed to be symmetrical [13].

Her facial feature proportions determined according to principles of beauty theory used in cosmetic surgery, such as the ideal angle and ratios for the so-called "Venus line", or Baum ratio, defining the angle of projection of the nose, and the "1/3 rule" specifying equal vertical spacing between the chin, nose, eyebrows, and hairline [14].

2) Joints and Actuation

In the current version of ERICA, the torso and face can be controlled, but the arms and legs are not yet actuated. In total, ERICA's body has 44 degrees of freedom (DOF), depicted in Fig. 2, of which 19 are controllable.

Most joint control is focused on facial actuation for expressions and speech. The eyes have 3 DOF's and can be controlled synchronously in yaw, pitch, and convergence. Upper and lower eyelids and inner and outer eyebrows provide 4 DOF's. Another 4 DOF's are focused on mouth height, width, and upper and lower corners of the mouth, while the final 2 DOF's in the face actuate the tongue and the jaw.

The skeletal body axes shown in black in Fig. 2 (right) are actuated. Besides the eyes, these axes provide 6 independent

[1] https://www.aldebaran.com/en/humanoid-robot/nao-robot
[2] https://www.aldebaran.com/en/a-robots/who-is-pepper
[3] http://www.a-lab-japan.co.jp/en/

DOF's: waist yaw, waist pitch, synchronous vertical shoulder movement, neck pitch, neck yaw, and neck roll. The 30 joints drawn in white in Fig. 2 are passive. Future versions of ERICA will provide actuation for more of these joints.

All motion is generated by pneumatic actuators regulated by servo valves. The motor controller sends and receives data over a serial connection at a rate of 20 Hz. Unlike robots such as the Geminoid HI-2, in which external control boxes held all servo valve actuators, ERICA's servo valves are internal, resulting in a compact and portable design. Their operation is quiet enough to be inaudible during normal operation.

B. *Speech Synthesis*

The quality and expressive range of speech synthesis may strongly affect people's impressions of a robot, so a speech synthesis system should have the ability to convey nuanced vocal expressions. ERICA's speech synthesis is performed using a custom voice designed for Hoya's VoiceText software[4]. Default rendering of most sentences is typically smooth with intonation determined by grammar, and manual specification of pitch, speed, and intensity is possible.

In order to synthesize additional nuances of natural speech, we have introduced over 500 special tokens using markup tags that are used to express para-linguistic and emotional information, such as: agreement, admiration, interest, surprise, lack of sympathy, dissatisfaction, confusion, a sense that the robot has been convinced, or many other moods and emotions [15]. Many tokens are also focused on interjections and backchannel utterances, and a variety of tokens for laughs and gasps are also included. These were recorded separately with specific prosodic patterns in accordance with the tags.

The generated audio signal from the **speech synthesizer** is sent back to the robot to generate lip sync and body rhythm behaviors, as shown in Fig. 3 and explained in Sec II-D-2.

C. *Sensing*

Fitting the sensors necessary for natural communication into the mechanical limitations of the human form provides significant difficulty, so ERICA currently uses external sensors on a wired network for human position tracking, sound source localization, and recognition of speech and prosodic information. The elements of the sensing framework are shown on the left side of Fig. 3.

1) *On-board sensors*

ERICA's on-board sensing includes two 1280x1024 pixel 30fps NCM13-J USB cameras mounted in her eyes. Our platform uses OpenCV to perform **face tracking** on the video feeds from these cameras to generate data that can be used for visual servo control of gaze.

Two Sony ECM-C10 omnidirectional condenser microphones (8mm diameter x 18mm length) are embedded in the ears. Although we hope to eventually use these microphones for speech recognition, it may be some time before this is practical. However, they can still be used for detection and coarse localization of sound activity.

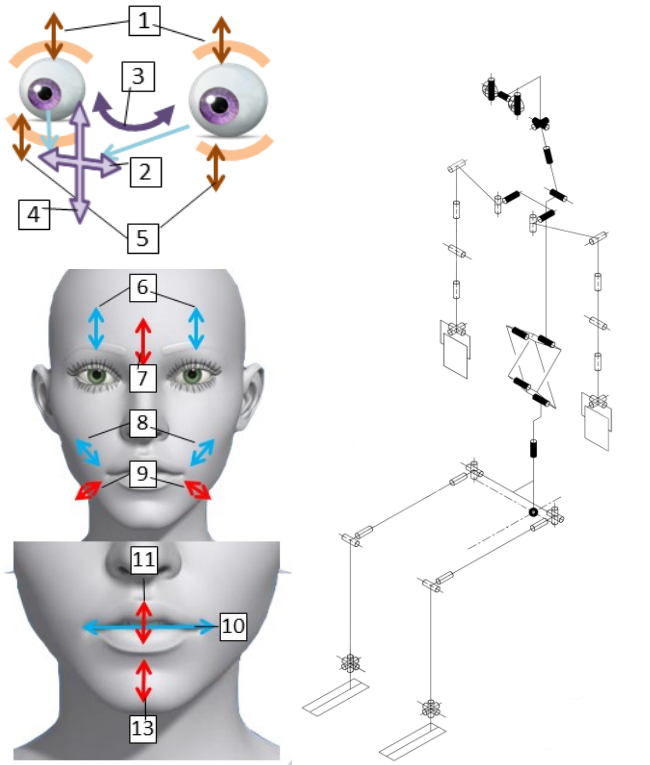

Figure 2. Degrees of freedom in ERICA. Left: Facial degrees of freedom. Right: Skeletal degrees of freedom. Joints marked in black are active joints, and joints drawn in white are passive.

2) *Position Tracking*

For ERICA, robust human **position tracking** is important for two reasons: first, for precise gaze control, and second, for keeping track of the identities of people (who may be moving) in multiparty interactions.

ERICA uses the ATRacker tracking system[5], which can be used with Microsoft Kinect 2 sensors, 2D laser range finders [16], or a network of ceiling-mounted 3D range sensors, presented in [17]. Each of these configurations has different advantages such as precision, portability, or scalability.

Accuracy varies depending on the types, number, and placement of sensors, but some studies have reported 11cm accuracy for the 2D tracking technique [18] and 17cm accuracy for the ceiling-mounted 3D tracking technique [19]. We are currently investigating whether additional techniques may be necessary for convincingly maintaining eye contact.

3) *Sound Source Localization*

In multiparty scenarios, it is important to know not only what was said, but also who is speaking. We perform **sound source localization** using two microphone arrays to detect the DOA (direction of arrival) of speech. The direction estimates from the microphone arrays are then combined with tracking data from the human tracking system [20, 21].

ERICA uses two 16-channel microphone arrays, from which sound directions are estimated in 3D space (azimuth and elevation angles) with 1 degree angular resolution and 100ms time resolution. If detections by multiple arrays

[4] http://voicetext.jp/

[5] http://www.atr-p.com/products/HumanTracker.html

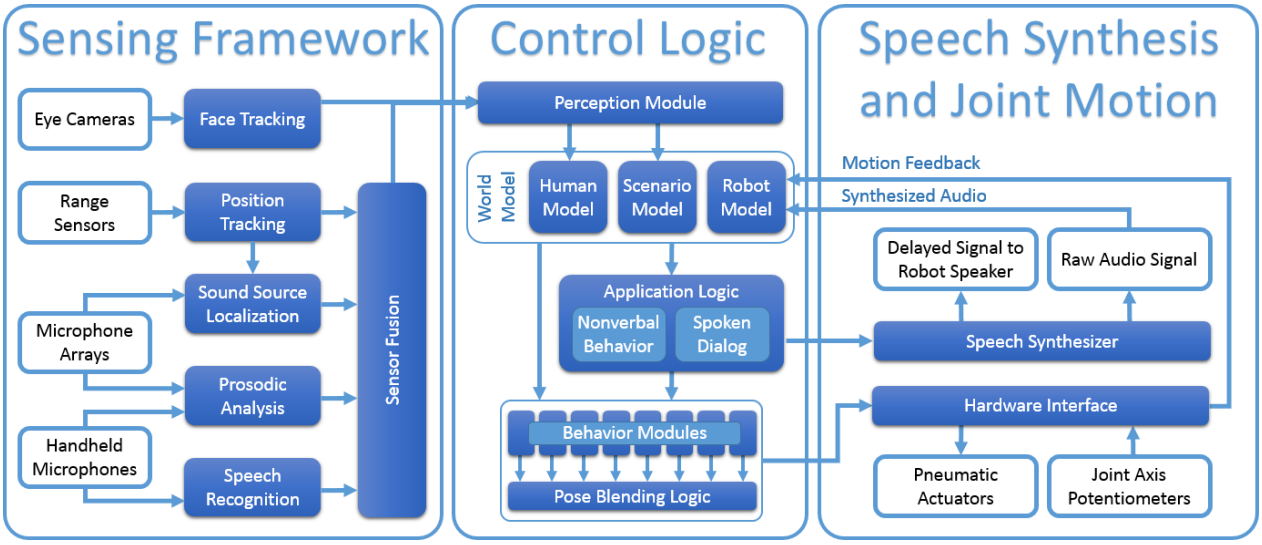

Figure 3. System diagram illustrating sensor inputs, internal control logic, and interaction with speech synthesis and motion generation.

intersect in 3D space and a human is tracked at that position, it is likely that that person is speaking.

Voice activity information can then be determined for each human around the robot, based on sound directivity. This is more robust than classical methods based on single channel acoustic power information, Furthermore, an estimate of the mouth height of the human is also provided, which can be used to control the gaze of the robot.

4) Speech Recognition

Speech recognition technologies have become widely available, especially for smartphones. However, many systems assume close-talk microphones and simple sentences or keywords. Speech recognition using microphones placed some distance from the speaker poses significant challenges. Typically, it requires calibration of many parameters based on measurement of the room acoustics.

For this reason, we are currently using wireless Shure Beta 58A handheld microphones for speech recognition. However, one of the research goals of this project is to develop the appropriate models and techniques for achieving reliable speech recognition in social conversation using microphone arrays placed at a distance. This will be realized by using deep neural networks (DNN) for front-end speech enhancement and acoustic modeling [22].

Speech recognition for Japanese is performed using the DNN version of the open-source Julius large-vocabulary speech recognition system [23]. Speech recognition for other languages is currently under development.

5) Prosodic Information

Prosodic features such as power and pitch of a person's voice play an important role in turn-taking, for signaling emotion or uncertainty, and for signaling questions or statements. An android will need to understand such signals in order to react quickly to the dynamics of a conversation, perform backchannel behaviors, change its gaze target, or express reactions before the human has finished speaking.

To enable such behaviors, our system analyzes the audio streams from speech inputs and provides continuous estimates of the power and pitch of the speaker's voice, as well as an estimate of whether the current signal represents speech activity or not [15]. This analysis is done both for the close-talk microphone inputs and for the separated signals from the microphone arrays.

D. Control Architecture

The software architecture of the ERICA platform combines a memory model, a set of behavior modules for generating dynamic movements, and a flexible software infrastructure supporting dialog management. The center area of Fig. 3 illustrates the core elements of the interaction logic.

1) Perception and Memory

The robot's awareness and memory of the world are stored in what we call the **world model**, which is divided into a set of human models, a scenario model, and a robot model.

A **human model** contains information about a person's location, speech activity, recognized speech results, and dialogue-related content such as name, preferences, goals, or elements of interaction history. The **scenario model** holds information that is part of a social scenario. For example, if the robot and humans are engaged in a game, information about the state of the game would be held in the scenario model. Finally, the **robot model** includes a kinematic model of the robot's pose as well as current values of pitch and voice power computed from the output of the speech synthesizer.

2) Motion Control

The bottom of the center panel of Fig. 3 depicts a set of several **behavior modules** running in parallel, the output of which is combined in the **pose blending logic** component. Behavior modules can be activated or deactivated according to an end user's application logic.

Several behavior modules generate motion based on speech activity, including a "lip sync" behavior module, which calculates mouth and jaw commands to send to the robot based on the raw audio signal from the **speech synthesizer** [24]. Furthermore, while the robot is speaking, a "rhythm" module generates forward and backward motions of the body trunk based on the power and pitch of the robot's

TABLE I. BEHAVIOR MODULES

Pose and Animation	
Expressions	Manages transitions between fixed facial expressions and body poses
Gestures	Can execute one or more gesture animations in parallel
Idle Motions	
Breathing	Moves shoulders and torso to simulate breathing when the robot is not speaking
Blinking	Blinks eyes at a fixed rate with slight random variation
Speech-related	
Lip Sync	Moves robot's lips to match the robot's speech signal
Rhythm	Generates emphasis and nodding behaviors in response to the robot's speech signal
Backchannel	Generates nodding behaviors in response to the human's speech signal
Gaze Control	
Gaze Controller	Uses closed-loop control to manipulate 7 joints to direct the robot's gaze at a specific direction or point in space
Gaze Avert	Adds small offsets to the target gaze direction to simulate natural human gaze variations

speech signal [25]. While the robot is listening, a "backchannel" module is activated, which produces nodding behaviors based on short pauses in the person's speech. To provide time for motor actuation of the lip sync and rhythm modules, an empirically determined delay of 200 ms is applied to the output audio signal played through the robot's speaker.

Several other behavior modules have also been implemented, and a summary is presented in Table I. We plan to create a core set of behavior modules which will be useful in most interactions, although end users may wish to create some additional behavior modules for specific scenarios.

3) Application Logic

Every application of an android will have different scenario-specific requirements. The **application logic** must handle the high-level control of the android's attention, **nonverbal behaviors**, and **spoken dialog** logic, as well as factors such as the robot's emotion and attitude.

For example, the android will need to discriminate between the different roles of the people it will interact with. In the public demonstration described in Sec. VI, the application logic managed attention and dialog separately for people in different roles – ERICA's behaviors towards visitors asking questions were different from her behaviors toward the researchers answering questions beside her.

4) Dialog Management

Many conventional dialog management techniques for humanoid robots can be roughly categorized into *robot-initiative*, in which the robot mainly talks based on a given scenario, and *user-initiative*, typical in smartphone assistant systems such as Apple's virtual assistant Siri, in which the robot only responds to user queries. This means the system does not act unless the user says something. Our goal is to realize natural *mixed-initiative* dialog, in which the robot both responds to user queries and takes its own initiative.

The system is currently using a state-transition model for dialog control. The dialog module compares received speech recognition results against a list of keywords. The process is relaxed from exact matching to statistical matching, and also from the first-best hypothesis to the fifth-best hypothesis. The matching threshold is set high to prevent false matches.

In the case of a match, an internal state machine is updated and an utterance is generated for the robot to speak. When no successful match can be made, we try to mimic human behaviors using facial expressions and short reactive tokens such as "hmm?" rather than repeating phrases like, "I'm sorry, I cannot understand you." Each output utterance may be combined with a facial expression and/or gesture.

In this way, the android can both answer the user's queries and also ask questions back to the user and express a reaction to the user's reply. Although this mechanism is quite simple, it is able to handle multiple-turn conversations and incorporate history, which is enough to successfully address people's questions and ask questions back to those people.

III. APPLICATION DEVELOPMENT

To illustrate an example of the development process for an autonomous android application, we will use the example of ERICA's first public demonstration in "Miraikan", the Japanese National Museum of Emerging Science and Innovation in Tokyo. The target of this demonstration was to showcase the abilities of ERICA by having her introduce herself and answer questions from the press. Furthermore, we wanted to demonstrate multiparty interaction by having her interact with the researchers who were on stage alongside her.

A. Dialog Mechanism

To enable multiparty interaction, the dialog management system was configured to support parallel state machines with independent sets of rules. In practice, we used two state machines: one for questions from members of the audience, and one for utterances from the other people on stage. Multiple visitors could be on stage at once, but only one could speak at a time, since there was only one visitor microphone.

The sensing system distinguished between utterances from different individuals based on which microphone was being used. When someone spoke, the speech recognition data received from their wireless microphone was fused with the human position data from the LRF-based tracking system by using sound source localization from the microphone arrays. In this way, the system was able to identify the identity (role) and location of the speaker, and to track them as they moved.

When a speech event was received, it was sent to the appropriate dialog manager based on the identity of the speaker. If the dialog manager judged that ERICA should respond to the speaker, that person was then set as her attention target for gaze and backchannel behaviors, and the robot then executed a sequence of utterances, gestures, and facial expressions determined by the dialog manager.

B. Conversation Content

The utterance content and transition rules for this demonstration were all scripted by hand over a period of about two weeks. Interaction content was prepared for 30 question topics that a visitor could ask and 40 comments that the researchers could make. To recognize the possible variations of each speech input, approximately five alternative keywords were designated for each topic. ERICA also kept track of the

dialog history and varied her responses when a question was asked multiple times.

In total, 731 utterances were prepared for ERICA to speak. Only 353 of these utterances were textual phrases or sentences. The remaining 378 utterances were interjections or backchannel utterances (in Japanese, these are sounds like "un" or "he---"), scripted with subtle differences in vocal inflection and emotional expression. The designer tested each utterance and applied markup tags to adjust the nuances of the pronunciation. In total, 993 speech tag instances were used to control pitch, speed, emotion, or pauses in the spoken text.

C. Nonverbal Behavior

29 facial expressions and 16 gesture animations were developed for this demonstration. These included motions like bowing and nodding, as well as expressions of joy, sadness, surprise, disappointment, indecision, doubt, relief, wry humor, dissatisfaction, and many others.

Together with these explicit expressions and gestures, the nonverbal behaviors specified in Table I were used to generate breathing, blinking, gaze, backchannel, and other "implicit" behaviors. A minor bug (fixed after the demonstration) caused the lip sync behaviors to lag somewhat, but other behaviors worked as designed. These allowed ERICA to produce a variety of nonverbal behaviors during the demonstration. For example, when no visitor was detected, she looked randomly out at the crowd, periodically giving a small smile to people. She looked at the visitor if one was present, and when the visitor spoke, she performed backchannel nodding to signal understanding during the short pauses in the visitor's speech.

D. Gaze control

As ERICA needed to interact at different times with four different people, it was necessary to implement logic to manage her attention target, primarily for controlling her gaze and backchannel behaviors. When one of the researchers or the MC spoke, she glanced at that person for a few seconds, then looked back at either the visitor or the crowd. When she spoke, she looked in the direction of the addressee (the visitor or a researcher when she was answering a question, or the crowd when she was giving a self-introduction). This logic was based on context and social rules, so it was hand-coded for each phase of the demonstration (self-introduction phase, researcher presentation phase, question-and-answer phase, etc.). These phases were based on the day's event schedule and manually started by a human operator.

E. Interaction Design Strategy

To achieve an appearance of humanlikeness, we believe it is important for a robot to convey a sense that it has its own desires and intentions. Following this principle, we created simple mixed-initiative interactions. For each topic, after ERICA answered a question, she continued by asking a question back to that person. For example, if a reporter asked about her hobbies, she would answer the question and then ask about the reporter's hobbies. She would then make a short comment on the reporter's answer. We limited interactions to two conversational turns (human-robot-human-robot) in order to limit the complexity of the dialog model.

Another strategy for conveying a sense that the robot had its own desires and intentions was to create utterances which were emotionally expressive or judgmental, such as being disappointed in answers that she didn't like or being offended when a very personal question was asked. We believe that interactions like these, which are mixed-initiative and in which the robot expresses opinions and emotions, can create a sense that the robot has some level of intentionality.

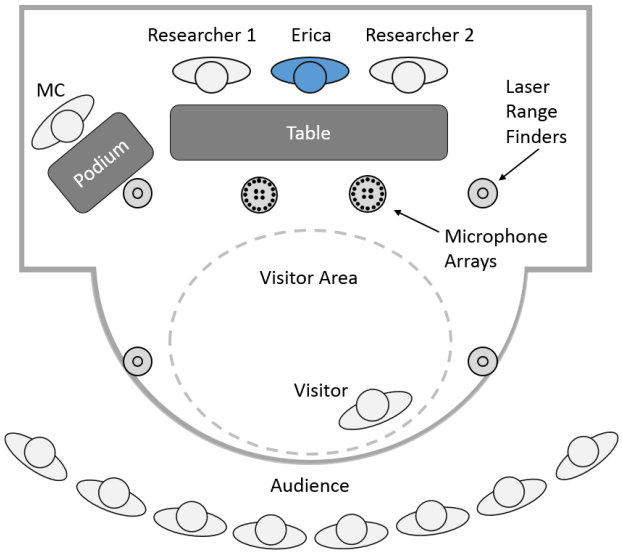

Figure 4. Stage arrangement for the public demonstration.

IV. Interactive Demonstration

In the actual demonstration, ERICA was seated behind a table on a stage, along with two researchers who were answering questions about the project. The center of the stage was monitored with two microphone arrays and four planar laser range finders, as shown in Fig. 4. Members of the press and the public were invited to come on stage and ask questions to ERICA or the researchers using a wireless microphone, as shown in the photo in Fig. 5.

A list of 30 topics were shown on a projection screen, and visitors took turns asking ERICA about those topics. After responding to each question, ERICA asked a question in return, based on the dialog state history. For example (translated from Japanese):

Visitor: *How old are you?*

ERICA: *I'm 23 years old. Even though I was just built, please don't call me 0 years old. (laughs)*

ERICA: *Do you think I look older?*

Visitor: *Yes, I think so.*

ERICA: *(giggles and smiles) Thanks! People always think I look younger, so I'm happy to hear that.*

ERICA also responded to utterances of the researchers and the MC at different times in the demonstration. The visitor, the MC, and the two researchers each had separate microphones, and each microphone was independently processed for speech recognition and prosodic information. This enabled ERICA to respond to each person in an appropriate way. For example:

Researcher: *(Turns to ERICA after answering a visitor's question). ERICA, you're the greatest robot ever, aren't you?*

ERICA: *(Turns to the researcher and smiles) Yes! (Then, after a short pause, makes a worried expression) Well… actually, we'll see. That depends on how well my researchers program me.*

Overall, we considered the demonstration to be a great success. It received many positive reviews from the press, and we encountered no significant technical problems.

V. Achievements and Future Work

This demonstration of the ERICA platform marked an important milestone in this project and showcased ERICA's basic capabilities. Although we have not conducted formal evaluations of the robot, we will discuss in this section our qualitative impressions what we have achieved so far and what future work lies ahead.

A. Hardware Platform

It is difficult to quantify what has been achieved in terms of the robot's physical appearance, but, anecdotally speaking, it is encouraging that at least one news agency[6] reported on the demonstration with the headline, "Japan's Erica android isn't as creepy as other talking robots." Given that popular media has a tendency to focus on the uncanniness of humanlike androids, we interpreted this as a sign of good progress.

For this demonstration, the only actuators available were for Erica's waist, neck, shoulders, and face. In the future, full-body poses and expressivity will be necessary. As this project continues, we plan to add articulation to the arms and create a mobile version of the robot.

B. Speech Synthesis

We were quite satisfied by the naturalness and expressivity of the speech synthesis. However, each of the speech tags had to be manually inserted by the author of the utterance content, making the process of content creation quite painstaking and slow. The design process could be accelerated if some of these tags could be assigned in formulaic ways, for example, by specifying only an abstract mood or feeling to apply to an utterance, and automatically assigning not only speech tags, but also gestures or expressions to that utterance.

C. Nonverbal behavior

1) Explicit expressions and gestures

Most of the facial expressions created were quite subtle, and occasionally the nuances were hard to distinguish. However, it also seemed to us that these nuanced expressions made the robot seem more humanlike than the exaggerated expressions made by some robots and virtual agents. With ERICA's hardware configuration, it would be difficult to create very dramatic expressions, but for everyday tasks, subtle expressions would likely be more useful, especially given the modest level of expressivity in Japanese culture.

2) Implicit behaviors

During ERICA's interactions, implicit behavior modules were used to actuate breathing, blinking, gaze, speaking rhythm, and backchannel nodding. We found these behaviors to be very useful in creating a feeling of lifelikeness in the robot. Because the robot could automatically generate motion based on audio signals and sensor inputs, these implicit behaviors enabled a high degree of lifelike movement while reducing the burden of the interaction designer to explicitly program motion commands and gestures.

Figure 5. Photo of the public demonstration.

In the future, we hope to improve and formalize each of these modules and develop a variety of new implicit behaviors, such as motion control for laughter, unconscious fidgeting, and methods of expressing emotion implicitly through adjustments of gaze and body movement.

D. Multimodal Perception

We considered the capabilities of ERICA's sensor network to have been quite sufficient for this demonstration. However, to achieve deeper and more engaging interactions, it will be important to react to nonverbal cues from people as well.

For that purpose, we plan to incorporate information from RGB-D cameras to understand the gestures and expressions of people around the robot. Paralinguistic information conveyed by speech will also be collected, by accounting for prosodic information extraction in noisy environments.

E. Desire and Intention

One of the most challenging questions regarding autonomous androids is how to design a robot's "mind". Currently, ERICA's application logic is all manually crafted as sequences of utterances. Many other robot systems are developed this way, and we plan to incorporate visual tools such as Interaction Composer [26] to assist the process of interaction design. We also plan to develop ways to train the robot from example interaction data, both through teleoperation [27], and through passive observation of human-human interaction [28].

Eventually it will be necessary to generate behavior based on representations of semantic meaning and desire and intention of the robot. We plan to investigate ways to implement a hierarchical model consisting of high-level "desires" which generate low-level "intentions", from which specific "behaviors" can be generated. Desire and intention are not merely tools for creating humanlikeness; they also form a basis for purposeful planning and a mechanism for the programming of high-level goals of an autonomous android.

VI. Conclusion

In this paper we have presented ERICA, a novel android

[6] http://mashable.com/2015/08/12/erica-android-japan/
This page also links to a video showing the android demonstration.

platform created for research into android autonomy. We have explained the principles and implementations of ERICA's hardware and software architecture and described the development process of a simple android interaction. We hope that the android community finds this useful, and we believe that it contributes to a larger discussion of what design methodologies will be effective for android systems.

We believe ERICA is the most humanlike android today, thanks to her visual design, facial expressivity, and highly expressive speech synthesizer. Her sensing technologies are some of the most capable to date, with high-performance speech recognition, the ability to discriminate between sound sources using microphone arrays, and precise tracking of people's locations and movements. This work thus helps us understand what is possible given the current state of the art, and it helps us to drive out the key issues and understand the next steps on the path to creating truly humanlike androids.

The ultimate goal of this research is to create an android which can communicate in a convincingly humanlike way in face-to-face interactions. While this problem is sort of a "grand challenge" for android science, we believe it can be achieved in the near future for focused, specific scenarios. We have outlined the key challenges that we plan to address in order to achieve this goal, and we look forward to reporting on advances in those areas as we work towards greater levels of autonomy, humanlikeness, and interactive capabilities.

ACKNOWLEDGMENT

We would like to thank Jani Even, Florent Ferreri, Koji Inoue, and Kurima Sakai for their contributions to ERICA's control software, dialog system, and sensor network.

REFERENCES

[1] K. Ogawa, S. Nishio, T. Minato, and H. Ishiguro, "Android Robots as Tele-presence Media," Biomedical Engineering and Cognitive Neuroscience for Healthcare: Interdisciplinary Applications, pp. 54-63, 2012.

[2] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," Robotics & Automation Magazine, IEEE, vol. 19, pp. 98-100, 2012.

[3] A. Hartholt, D. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch, "All together now: Introducing the Virtual Human Toolkit," in Intelligent Virtual Agents, 2013, pp. 368-381.

[4] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in Cognitive Behavioural Systems, ed: Springer, 2012, pp. 114-130.

[5] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda, "Humanoid robots as cooperative partners for people," Int. Journal of Humanoid Robots, vol. 1, pp. 1-34, 2004.

[6] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Development and evaluation of interactive humanoid robots," Proceedings of the IEEE, vol. 92, pp. 1839-1850, 2004.

[7] C. Diana and A. L. Thomaz, "The shape of Simon: creative design of a humanoid robot shell," presented at the CHI '11 Extended Abstracts on Human Factors in Computing Systems, Vancouver, BC, Canada, 2011.

[8] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou, "Upending the uncanny valley," in Proceedings of the national conference on artificial intelligence, 2005, p. 1728.

[9] S. Nishio, H. Ishiguro, and N. Hagita, Geminoid: Teleoperated android of an existing person: INTECH Open Access Publisher Vienna, 2007.

[10] C. Becker-Asano and H. Ishiguro, "Evaluating facial displays of emotion for the android robot Geminoid F," in Affective Computational Intelligence (WACI), 2011 IEEE Workshop on, 2011, pp. 1-8.

[11] K. Ogawa, S. Nishio, K. Koda, G. Balistreri, T. Watanabe, and H. Ishiguro, "Exploring the Natural Reaction of Young and Aged Person with Telenoid in a Real World," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 15, pp. 592-597, 2011.

[12] H. Sumioka, S. Nishio, T. Minato, R. Yamazaki, and H. Ishiguro, "Minimal human design approach for sonzai-kan media: Investigation of a feeling of human presence," Cognitive computation, vol. 6, pp. 760-774, 2014.

[13] D. I. Perrett, D. M. Burt, I. S. Penton-Voak, K. J. Lee, D. A. Rowland, and R. Edwards, "Symmetry and human facial attractiveness," Evolution and human behavior, vol. 20, pp. 295-307, 1999.

[14] P. M. Prendergast, "Facial proportions," in Advanced Surgical Facial Rejuvenation, ed: Springer, 2012, pp. 15-22.

[15] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," Speech communication, vol. 50, pp. 531-543, 2008.

[16] D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita, "Laser-Based Tracking of Human Position and Orientation Using Parametric Shape Modeling," Advanced Robotics, vol. 23, pp. 405-428, 2009.

[17] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person Tracking in Large Public Spaces Using 3-D Range Sensors," Human-Machine Systems, IEEE Transactions on, vol. 43, pp. 522-534, 2013.

[18] D. F. Glas, F. Ferreri, T. Miyashita, H. Ishiguro, and N. Hagita, "Automatic calibration of laser range finder positions for pedestrian tracking based on social group detections," Advanced Robotics, 2012.

[19] D. F. Glas, D. Brščić, T. Miyashita, and N. Hagita, "SNAPCAT-3D: Calibrating networks of 3D range sensors for pedestrian tracking," in Robotics and Automation (ICRA), 2015 IEEE International Conference on, 2015, pp. 712-719.

[20] C. T. Ishi, J. Even, and N. Hagita, "Integration of Multiple Microphone Arrays and Use of Sound Reflections for 3D Localization of Sound Sources," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 97, pp. 1867-1874, 2014.

[21] C. T. Ishi, J. Even, and N. Hagita, "Speech activity detection and face orientation estimation using multiple microphone arrays and human position information," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015), 2015, p. (Accepted for publication).

[22] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," EURASIP Journal on Advances in Signal Processing, vol. 2015, pp. 1-13, 2015.

[23] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, 2009, pp. 131-137.

[24] C. T. Ishi, C. Liu, H. Ishiguro, and N. Hagita, "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," in Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, 2012, pp. 2377-2382.

[25] K. Sakai, T. Minato, C. T. Ishi, and H. Ishiguro, "Speech Driven Trunk Motion Generating System Based on Physical Constraint," presented at the 43rd Japanese Society for Artificial Intelligence AI Challenge, Keio University, Kanagawa, Japan, 2015.

[26] D. F. Glas, S. Satake, T. Kanda, and N. Hagita, "An Interaction Design Framework for Social Robots," in Proceedings of Robotics: Science and Systems, Los Angeles, CA, USA, 2011.

[27] K. Wada, D. Glas, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Capturing Expertise: Developing Interaction Content for a Robot Through Teleoperation by Domain Experts," International Journal of Social Robotics, pp. 1-20, 2015/02/27 2015.

[28] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to train your robot - teaching service robots to reproduce human social behavior," in Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on, 2014, pp. 961-968.