Evaluation of Head Gaze Loosely Synchronized With Real-Time Synthetic Speech for Social Robots

Vasant Srinivasan, Cindy L. Bethel, and Robin R. Murphy, *Fellow, IEEE*

***Abstract*—This study demonstrates that robots can achieve socially acceptable interactions using loosely synchronized head gaze-speech acts. Prior approaches use tightly synchronized head gaze-speech, which requires significant human effort and time to manually annotate synchronization events in advance, restricts interactive dialog, or requires that the operator acts as a puppeteer. This paper describes how autonomous synchronization of head gaze can be achieved by exploiting affordances in the sentence structure and time delays. A 93-participant user study was conducted in a simulated disaster site. The rescue robot "Survivor Buddy" generated head gaze for a victim management scenario using a 911 dialog. The study used pre- and postinteraction questionnaires to compare the social acceptance level of loosely synchronized head gaze-speech against tightly synchronized head gaze-speech (manual annotation) and no head gaze-speech conditions. The results indicated that for attributes of Self-Assessment Manikin, i.e., Arousal, Robot Likeability, Human-Like Behavior, Understanding Robot Behavior, Gaze-Speech Synchronization, Looking at Objects at Appropriate Times, and Natural Movement, the loosely synchronized head gaze-speech is similar to tightly synchronized head gaze-speech and preferred to the no head gaze-speech case. This study contributes to a fundamental understanding of the role of social head gaze in social acceptance for human–machine interaction, how social gaze can be produced, and promotes practical implementation in social robots.**

***Index Terms*—Autonomous generation, human–robot interaction, interactive conversation, social head gaze, user study and evaluation.**

I. Introduction

A KEY component of social interaction between a robot and human(s) is social head gaze, which serves five important functions—*looking interested in humans* [1]–[15], *engaging in a verbal conversation* [4], [7]–[9], [11], [13]–[18], *exhibiting aliveness* [2], [6], *showing various mental states*[2], [19], and *referential gaze* to objects in the environment [2], [6], [13]–[15], [20]–[22]. Two of these functions, i.e., *engaging in a verbal conversation* and *referential gaze* to objects in the environment, require synchronization because communication occurs across two different but highly interdependent channels: head gaze and speech. While *engaging in a verbal conversation* with a human, the robot generates fixate and avert gaze acts that are tightly synchronized with speech to facilitate turn taking. If the topic of the discussion is an object in the environment, the robot uses *referential gaze* to fixate toward the object 800 ms to 1 s, before it utters the object's name. The tight synchronization between head gaze and speech (TSHG-S) has been well modeled in the human–human literature [23]–[26] and ensures high-quality communication between humans. However, social robots using models for turn taking in conversations [23], [24] and referential gaze for looking at objects in the environment [25], [26] suffer from three limitations. First, TSHG-S requires manual annotation and semantic content understanding. This requires significant human effort and time. Second, if the robot uses a preset library to select appropriate head gaze behaviors, the head gaze cannot be generated in open-ended interactive scenarios. This is problematic when the social robot's verbal responses cannot be anticipated *a priori*. Third, tight synchronization of gaze and speech that mimics human gaze may not be feasible due to limitations of the robot. These might include the absence of a high degree of motor control, flexibility of joint movements, and/or velocity limits.

This paper examines the use of what is known as *affordances* to generate *socially acceptable* head gaze acts loosely synchronized with real-time synthetic speech (LSHG-S) for human–robot interaction. In behavioral robotics, *affordances* are conditions or objects that are directly perceivable without any memory, inference, or interpretation [27]. Social acceptance is a measure of how well the robot performs across three categories of measures: *Participants' Positive Affective State, Participants' Perception of the Robot,* and *Consistency and Appropriateness of the Robot's Head Movements*. The occurrence of turn events and semantics in dialog that activate head gaze acts can be substituted with affordances from the sentence structure of dialog and time delays. These affordances are computationally trivial, support autonomous generation, are independent of semantics, and are useful for interactive open-ended conversations. If a robot is an autonomous agent and can generate dialog, the sentence structure and time delays will be transparently available to the robot, and the proposed method can be used. In the case of a teleoperated robot [28], [29] or wizard of oz experiment [30], the proposed method can be utilized if the robot operator can provide the dialog, from which the sentence structure and time delays can be determined.

Manuscript received September 1, 2013; revised July 3, 2014; accepted July 11, 2014. Date of publication August 15, 2014; date of current version November 12, 2014. This work was supported in part by NSF IIS-0905485 "The Social Medium is the Message," RESPOND-R mobile, distributed test instrument created through NSF Grant CNS-0923203, and Microsoft External Research. This paper was recommended by Associate Editor C. M. Lewis.

V. Srinivasan and R. R. Murphy are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: vasants@cse.tamu.edu; murphy@cse.tamu.edu).

C. L. Bethel is with the Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: cbethel@cse.msstate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2014.2342035

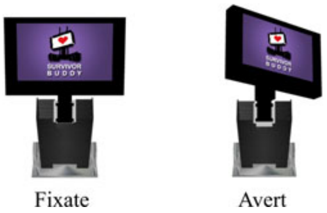

Fig. 1. Survivor Buddy *engaging in a verbal conversation* using fixate and avert gaze acts.

There are two potential problems with using LSHG-S. First, the loosely synchronized gaze acts may not precisely match the dialog presented, because there is no semantic understanding in the proposed approach. The lag between the robot's speech and gaze acts may annoy or confuse the human. Second, a social robot that interacts with a human in a realistic scenario will possibly use all five functions of head gaze. The change in synchronization for two functions such as *engaging in a verbal conversation* and *referential gaze* may impact the human's perception of the robot's overall head gaze during the interaction. Therefore, a study was conducted with the "Survivor Buddy" rescue robot (see Fig. 1) to determine whether LSHG-S is socially acceptable for human–robot interaction.

The scope of this study is limited to the investigation of head gaze generation in human–robot interaction, focusing on gaze-speech synchronization, and does not include generation of arm or manipulator gestures, such as pointing.

This paper is organized as follows. Section II provides a summary of related work. Section III describes the approach for the development of LSHG-S. Section IV discusses the implementation of LSHG-S, using affordances and production rules on the Survivor Buddy robot. Section V describes the methods of the human participant study. Section VI presents the data analysis and results from pre- and postinteraction questionnaires that demonstrate the social acceptance of LSHG-S. Section VII presents an interpretation of the results, discusses the factors that might influence the naturalness of head gaze-speech synchronization, and the limitations of the user study. Section VIII concludes that LSHG-S is adequate for human–robot interaction and ends with a short summary of the contributions of the study.

II. Related Work

Historically, social head gaze for *engaging in a verbal conversation* or *referential gaze* occurs across two different but highly interdependent channels: head gaze and speech. For example, in everyday face-to-face conversation, humans routinely use head gaze acts like fixate and avert in coordination with their speech. The remaining three functions: *looking interested in humans*, *exhibiting aliveness*, and *showing various mental states* do not use the speech channel, and hence, studies focusing on these functions are not considered in this review. Fourteen studies have been conducted on the topic to date [5]–[9], [11], [13]–[17], [19], [21], [22]. The criteria for examination of each study are *type of synchronization*, *synchronization method*, *type of speech*, *head gaze generation mechanism*, and the *human–human interaction model(s) used* (see Table I).

TABLE I
Tight Synchronization of Head Gaze With Speech in the Literature

Studies	Synchronization Method	Type of Speech	Gaze Generation Mechanism	Model
Matsusaka *et al.* [16]	Autonomous Identification of a Single Turn Event - End of Turn	Synthetic, Real Time	Production Rules	Turn Taking [24], [31]
Sidner *et al.* [6]	Manual Annotation of Turn Events	Synthetic, Real time	Production Rules	Turn Taking [23]
Macdorman *et al.* [5]	Manual Annotation	Human, Prerecorded	Prescripted Motion	Human–Human Experiment
Mutlu *et al.* [7]	Manual Annotation of Linguistic Events	Human, Prerecorded	Production Rules	Turn Taking [18]
Kuno *et al.* [8]	Manual Annotation of Turn Events	Synthetic, Human, Prerecorded	Production Rules	Turn Taking [24]
Yamazaki *et al.* [9]	Manual Annotation of Turn Events	Synthetic, Prerecorded	Production Rules	Turn Taking [24]
Staudte & Crocker [21], [22]	Manual Annotation of Linguistic Events	Synthetic, Prerecorded	Prescripted Motion	Referential Gaze [25], [26]
Mutlu *et al.* [11]	Manual Annotation of Turn Events and Footing Cues	Unspecified, Prerecorded	Production Rules	Human–Human Experiment, Turn Taking [23], [32], [33]
Mutlu *et al.* [19]	Manual Annotation of Linguistic Events	Human, Prerecorded	Prescripted Motion	Human–Human Experiment
Ishi *et al.* [17]	Manual Annotation of Linguistic Events	Human, Prerecorded	Production Rules	Human–Human Experiment
Holroyd *et al.* [13]	Manual Annotation of Turn Events Events	Synthetic, Real-Time	Production Rules	Turn Taking [23], [24]
Huang & Mutlu [14], [15]	Manual Annotation of Turn & Linguistic Events	Human, Pre-Recorded	Production Rules	Turn Taking [23], [24], Referential Gaze [25], [26]

Each of the 14 studies [5]–[9], [11], [13]–[17], [19], [21], [22] use TSHG-S. This method is precise and implements models of human–human interaction for conversational turn taking [23], [24], [31] and referential head gaze [25], [26], or relies on a human–human interaction experiment conducted specifically for determining the synchronization event [11], [17], [19]. TSHG-S has been shown to have many benefits for humans, such as increased task performance [6], [7], [11], [19], increased engagement [5], [6], [13], [19], improved understandability [6], [13], improved likeability [7], [11], [19], and increased positive feelings [6], [34]. However, to date, the use of LSHG-S has not been investigated.

Manual annotation is the most popular method for tightly synchronizing head gaze with speech. There are six different synchronization events discussed in the literature relating to manual annotation. The first three consist of turn events: *Start of Turn* [7]–[9], [13], [14], [16], [18], *Middle of Turn* [7], [18], and *End of Turn* [7]–[9], [13], [14], [18]. The last three include linguistic events: *First Word in Rheme* [7], [18], *First Word in Theme* [7], [18], and *Utterance of Object* [14], [15], [21], [22]. The theme specifies the topic of a sentence, i.e., what the sentence is all about, while the rheme specifies what is new or interesting about the topic [35]. A person identifies and marks the synchronization events in prerecorded audio files containing dialog [5], [7]–[9], [11], [13]–[15], [17], [19], [21], [22] or text used to generate synthetic speech [6], [13]. Manual annotation requires significant human time and effort. It requires that the robot head gaze behaviors be prescripted [19], [21], [22], or selected from a preset library of one-directional sentences and phrases using production rules [5]–[9], [11], [13]–[15], [17]. The use of prerecorded audio limits the interactivity of the dialog to the extent of the preset library; therefore, the robot is not able to adapt its dialog to the needs of the current interaction. Synthetic speech is advantageous for interactive conversations because speech can be generated in real time based on a textual input, and the robot can adapt to situations through the generation of dynamic dialog. However, without any methods that support autonomous annotation, the original limitations on interactivity still persist.

The only research that uses autonomous head gaze for conversation is by Matsusaka *et al.* [16], which models conversational strategy for robot participation in a group conversation. However, of the three possible turn events in conversations (*Start of Turn*, *Middle of Turn*, and *End of Turn*), this study only identifies autonomously the start of a turn. The implementation does not use any linguistic events and has not been validated in an experiment.

Findings from Kirchhof & Ruiter [36] suggest that gesture comprehension is temporally and semantically more flexible than gesture production in humans, and that a higher tolerance exists for modeling gestures in robots. The implementations of social head gaze are still at the manual Wizard-of-Oz stage, and no work has considered formal methods that transfer findings to support autonomous implementations. This paper examines the effect of LSHG-S, using close approximations of turn and linguistic events from human–human interaction models, to determine if it is adequate for human–robot interaction.

III. Approach

The approach toward the development of LSHG-S requires two steps: 1) the substitution of affordances for linguistic and internal contexts of head gaze, and 2) the use of production rules to map affordances onto gaze acts. Affordances are popular with behavioral roboticists, because they simplify computation [27]. For example, during a human–human conversation, a person always averts his or her gaze at the beginning of a turn, and at the start of the "theme" [18]. The limitation of using "theme" and "rheme" is that they are subjective and vary with sentences. A word that is at the beginning of the "rheme" in one sentence, need not mark the beginning of the "rheme" in another sentence. However, in straightforward simple sentences, the theme is at the beginning of the sentence, and the rheme is at the end [35]. Thus, punctuation (*! ?, and carriage return at the end of a paragraph*) in text-to-speech or inflection in voice recognition are approximations of the location of the "theme" or "rheme" and act as affordances. The advantages of using these affordances over the "theme" or "rheme" are that they are unique, can be easily computed, and support autonomous annotation.

Production rules are IF–THEN statements; if an affordance is perceived, the gaze act is called. Using production rules, a text sentence can be parsed, and the head gaze act(s) generated as the text is converted to speech. The production rules directly correspond to the five functions of head gaze, and a single head gaze function may be comprised of one or more production rules. The production rules are specific to the affordances observed, the gaze acts that are implemented, and the mechanical properties of the robot.

The proposed approach for the realization of head gaze through the use of affordances and production rules does not require manual annotation. It enables interactive open-ended conversations and can be tailored to the limitations of specific robots and applications.

IV. Implementation

The proposed method for loosely synchronizing head gaze with real-time synthetic speech was implemented on an affective robot for victim management, named "Survivor Buddy." The mechanisms for executing head gaze acts, perceiving the affordances, and the production rules used to map affordances on to gaze act(s) are discussed in this section. The experiment conducted to evaluate this implementation is discussed in Section V.

A. Robot Platform

The proposed head gaze generation method was implemented on Survivor Buddy, an affective multimedia head mounted on an Inuktun Extreme-VGTV robot (see Figs. 1 and 3). Survivor Buddy has four degrees of freedom (DOF) and is capable of very agile movements [37]. Survivor Buddy's head is a 7-in touch screen monitor with a webcam and microphone manufactured by MIMO Monitors. The neck of the robot contains the speaker system.

The software implementation of the system is in C# and uses Microsoft text-to-speech with the Microsoft Anna voice. The implementation currently requires typed text and punctuation. This is available transparently to a robot that is operated autonomously or can be provided by the operator if it is tele-operated.

B. Head Gaze Acts

The robot used five head gaze acts: *fixate, avert, concurrence, confusion, and scan*. The gaze acts are robot-dependent, as each robot would have its own implementation of acts based on its DOF and motor characteristics. The Survivor Buddy robot moved with an average velocity of 33 °/s for all gaze acts. The implementation specifics of the gaze acts matched the known parameters used by earlier implementations of head gaze in the literature and are described below.

1) *Fixate* moves the robot's head to a position facing the human directly. Fixation occurs for an indefinite duration until another gaze act activates [1]–[15].

TABLE II
AFFORDANCES FROM SENTENCE STRUCTURE AND TIME DELAYS

	Contexts from Literature	Substituted Affordances in Sentence Structure and Time Delays
1	First Word in Theme Start of Turn [7], [11], [18]	Initial Word
2	First Word in Theme Middle of Turn [7], [11], [18]	Word following Punctuation : . ! ?
3	First Word in Rheme Middle of Turn [7], [11], [18]	After 75% of Words between Punctuation : . ! ?
4	First Word in Rheme End of Turn [7], [11], [18]	Carriage Return
5	Internal State$_{\text{acknowledge}}$ [2], [6]	Elapsed Listening Time$>$ 6 s
6	Internal State$_{\text{aliveness}}$ [2], [6]	Elapsed Idle Time$>$ 15 s
7	Internal State$_{\text{confused}}$ [2]	Number of Deletes/Retypes by an Operator $>$ 5 within a Time Interval t = 15 s
8	800 ms to 1 s before Utterance of Object [14], [15], [21], [22]	The Object Name Tag

2) *Avert* is a +/− 7° simultaneous horizontal and vertical movement of the head, away from the fixation point. Aversion occurs for an indefinite duration until another gaze act activates [4], [7]–[9], [11], [13]–[18].
3) *Concurrence* is a repetitive vertical head movement of +/− 10° [6], [13]. Concurrence occurs once every 3 s.
4) *Confusion* is a head roll of +/− 20°. The head returns to the fixation point after 1 s [2].
5) *Scan* is a fixation persisting between 0.77 and 1 s to a series of three random points in space [2], [6], [11].

C. Affordances

The current implementation uses eight affordances *Initial Word*, *Word following Punctuation : . ! ?*, *After 75% of Words between Punctuation : . ! ?*, *Carriage Return*, *Elapsed Listening Time $>$ 6 s*, *Elapsed Idle Time $>$ 15 s*, *Number of Deletes/Retypes by an Operator $>$ 5 within a Time Interval t = 15 s*, and *the Object Name Tag*. These eight affordances from the structure of sentences used in dialog and time delays are used to infer contexts corresponding to synchronization events (e.g., *End of Turn*, *Start of Turn*, *Middle of Turn "Theme,"* and *Middle of Turn "Rheme"*) and internal states (e.g., *internal state*$_{\text{acknowledge}}$, *internal state*$_{\text{aliveness}}$, and *internal state*$_{\text{confused}}$) from the literature (see Table II). Table II lists the eight social contexts in the literature for turn events, linguistic events, and internal states, that require semantic or speech understanding (column 1) and how this research substitutes an affordance of either sentence structure or time delay (column 2) for a linguistic or internal context, which produces a social head gaze act.

The first four affordances (rows 1–4) in Table II are the sentence structure approximations of turn events (*Start of Turn*, *Middle of Turn*, and *End of Turn*) and linguistic events (*First Word in Rheme* and *First Word in Theme*) from models of human–human interaction. These approximations are: *Initial Word*, *Word following Punctuation : . ! ?*, *After 75% of Words between Punctuation : . ! ?*, and *Carriage Return*. The rationale for the approximation is that in the English language, the theme occurs at the beginning of an independent clause or simple sentence, and rheme occurs toward the end of the independent clause or simple sentence [35].

The next three affordances (rows 5–7) are approximations for the internal states of the robot. The affordance listed on row 5 approximates the *internal state*$_{\text{acknowledge}}$ by thresholding the time interval between the robot's responses to the human during dialog. If the *Elapsed Listening Time $>$ 6 s*, the internal state of the robot is set to acknowledge. The affordance for the *internal state*$_{\text{aliveness}}$ from row 6 is a timeout based on the idleness of the robot. If the *Elapsed Idle Time $>$ 15 s*, the internal state of the robot is set to aliveness. Since the existing literature does not provide guidance on specific values for back-channels and acknowledgements [6], [13], or aliveness [2], a suitable timeout was estimated by the researchers to communicate the corresponding function of head gaze effectively.

The affordance listed in Row 7 is based on typing and used only when an operator is using the robot. This affordance approximates the *internal state*$_{\text{confused}}$ [2] of the robot by thresholding the number of deletes and retypes in a time interval t. If the *Number of Deletes/Retypes by an Operator $>$ 5 within a Time Interval t = 15 s*, the internal state reflects confusion.

The affordance listed in Row 8 is a *Object Name Tag* inserted before an object name. This affordance approximates the gaze at an object in the environment from *800 ms to 1 s before Utterance of Object* [14], [15], [21], [22] to gaze at the object at the onset of Utterance of Object. The *Object Name Tag* can be inserted either manually when there is an operator present and typing a sentence or can be autonomously inserted by a reasoning system.

D. Production Rules

A total of nine production rules (see Table III) identified by the reference architecture for social head gaze [34], [38] were implemented on the Survivor Buddy robot to map the function on to the gaze act. Production rule 1 is used by the robot for *Looking Interested* in a human. The robot fixates toward the human to indicate attention [1]–[15], [39], [40]. Production Rules 2–6 are designed for *engaging in a verbal conversation* with the human. The robot uses rules of turn taking to fixate and avert from the human [4], [7]–[9], [11], [13]–[15], [17], [18]. *Exhibiting aliveness* is implemented on the robot using production rule 7. The robot uses the *scan* gaze act to randomly look at different points in space, and to indicate it is alive and functioning properly [2], [6]. *Showing Various Mental State*, such as confusion, is accomplished by production rule 8 [2]. Production Rule 9 is used to implement *Referential Gaze*, where the robot fixates toward an object in the environment when it utters the object name in speech [2], [6], [13]–[15], [20]–[22].

These nine production rules reflect the five functions of social head gaze and provide an exhaustive coverage of the social head gaze phenomena. The implementation of the production rules occur in parallel and use separate threads; therefore, multiple production rules may be active at any time t and need to be coordinated. Coordination ensures that synchronization occurs across two different channels—speech and gaze. It also ensures that the robot is sensitive to the current social function and conveys the appropriate meaning. Production rule coordination is accomplished in this implementation by allowing rules

TABLE III
NINE PRODUCTION RULES THAT MAP AFFORDANCES ONTO HEAD GAZE ACTS

	Function	Production Rule
1	Looking Interested in Human[1]–[15]	IF **Approach of Human**, THEN **Fixate** toward the human for an indefinite duration
2		IF **Initial Word**, THEN **Avert** from the human with a +/− 7° simultaneous horizontal and vertical movement for an indefinite duration
3		IF **Word following Punctuations . ? !**, THEN **Avert ($p = 0.73$)** from the human with a +/− 7° simultaneous horizontal and vertical movement for an indefinite duration
4	Engaging in a Verbal Conversation [4], [7]–[9], [16], [11], [13], [17], [18] [14], [15]	IF **After 75% of Words between Punctuation : . ! ?**, THEN **Fixate ($p = 0.7$)** toward the human for an indefinite duration
5		IF **Carriage Return**, THEN **Fixate** toward the human for an indefinite duration
6		IF **Elapsed Listening Time > 6 s**, THEN **Concurrence** toward the human with repetitive vertical head movement of +/− 10° every 3 s
7	Exhibiting Aliveness [2], [6]	IF **Elapsed Idle Time > 15 s**, THEN **Scan** three random points in the environment
8	Showing Various Mental States [2], [19]	IF **Number of Deletes/Retypes by an Operator > 5 within a Time Interval t = 15 s**, THEN **Confusion** toward the human with a head roll of +/− 20° and return to the fixation point
9	Referential Gaze [21] [2], [6], [13]–[15]	IF **The Object Name Tag**, THEN **Fixate** toward the object in the environment

LSHG-S	**TSHG-S**
<Avert, 1> You have been found in an area of the collapsed building that suffered <Fixate, .7> a lot of damage. <Avert, .73> Did you happen to see what caused <Fixate, .7> the collapse? <Fixate, 1>	<**Avert, 1**> **You** have been found in an area of the collapsed building *<Fixate, .7>* *that* suffered a lot <**Avert, .73**> of damage. **Did** you *<Fixate,1>* happen to *see* what caused the collapse?

Fig. 2. Annotated example from the script. The *First Word in Theme* is indicated by **Bold**, and the *First Word in Rheme* is indicated by *Italics*.

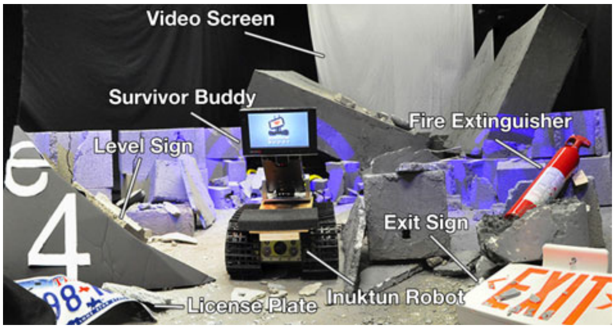

Fig. 3. Simulated disaster area from the participant's point of view.

8 and 9 to run to completion without interruption. All other production rules (1–7) can be interrupted by the most recent production rule. Some of the gaze acts are probabilistic in nature ($p =$ value) indicating the probability of activation of the gaze act. These production rules were identified from the literature (see Table III) of head gaze in human–robot interaction and can be expanded if new studies identify new uses of head gaze.

E. Differences Between LSHG-S and TSHG-S Implementations

The differences between the LSHG-S and TSHG-S implementations were in the timing and occurrence of head gaze acts in coordination with speech. This is illustrated with an example from the script: "You have been found in an area of the collapsed building that suffered a lot of damage. Did you happen to see what cause the collapse?"

The activation of production rules for LSHG-S is as follows. The input from the script fills the buffer variable that interleaves both speech and head gaze. The affordance *Initial word* "You" activates production rule 2 and inserts the *<Avert, 1>* head gaze tag. Next, affordance *After 75% of words between punctuations* activates production rule 4 and inserts *<Fixate, 0.70>* after the word "suffered." An *<Avert, .73>* gaze tag is inserted when the affordance *Word following a punctuation .* is perceived, before the word "Did" as described by production rule 3. Using production rule 4, another middle of turn *<Fixate, .7 >* is inserted after the word "caused" when the affordance *After 75% of words between punctuations . and ?* is perceived. Finally, the affordance *Carriage Return* activates production rule 5 to insert a *<Fixate, 1>* to indicate the end of turn. The contents of the buffer variable at the end of the robot's current turn is shown in Fig. 2. This passes through the Microsoft speech system, which converts the text-to-speech and triggers the head gaze act when it encounters a head gaze tag. The runtime for this implementation is $O(n)$, where n is the length of the text that needs to be parsed.

The head gaze generation for TSHG-S uses the contexts from the literature, such as *First Word in Theme*, *First Word in Rheme*, etc. (see Table II). These were identified by manual inspection using definitions described by Halliday [35] and marked with the corresponding head gaze acts on a prerecorded audio file (see Fig. 2). This same procedure was used in [7] and [18]. The Microsoft speech system triggers the head gaze act when it encounters a head gaze marker.

The timings for two head gaze acts—<Fixate, .7> and <Fixate, 1>—are different. In case of LSHG-S, <Fixate, .7> occurs after the word "suffered," whereas in TSHG-S, it occurs at the onset of the word "that." Similarly, the head gaze act <Fixate, 1> for LSHG-S occurs when a *Carriage Return* is perceived; however, for TSHG-S, that same <Fixate, 1> occurs at the onset of the word "see." Additionally, in the example described previously, LSHG-S had one extra <Fixate, .7> head gaze act after the word "caused."

V. METHODS

Three hypotheses concerning the impact of the LSHG-S were formed, and 23 measures were used to evaluate the overall social acceptance of the robot.

A. Experimental Design

A one-factor experiment was designed to evaluate three head gaze conditions: 1) LSHG-S; 2) TSHG-S; and 3) no head gaze-speech (NHG-S).

The scenario was a simulated disaster scenario of a parking garage collapse (see Fig. 3), wherein Survivor Buddy, the rescue robot shown in Fig. 3, has a dialog with a trapped victim based on 911 dispatch and triage protocols. This setup is an extension of [41], which illustrated that an extreme setting heightens affective responses from participants. The environment was comprised of the trapped victim, prop concrete floors, columns with rebar, simulated glass pieces, and objects typically found in a parking garage, such as a fire extinguisher, license plate, parking level sign, and exit sign. The environment also had a full theatrical stage lighting system in order to provide optimum visibility without sacrificing a lifelike effect. The participant interacted with Survivor Buddy for approximately 15 min. As per the 911 dispatch protocol, the dialog focused on assessing the participant's physical health and gaining information about the location and nature of the event. The dialog ensured the activation of each of the nine production rules at least once (none of the previous studies reported a dialog that captured all nine production rules) with a total of 162 possible gaze acts.

The experiment utilized *a priori* sensor data for object locations, participant head locations, and *internal state*$_{\text{confused}}$ of the robot. A hidden operator ("wizard") [30] stepped through conversation turns using predefined sentences and phrases rather than real-time typing. This approach overcame limitations in existing state-of-the-art object and speech recognition systems and ensured repeatability and consistency across conditions. This led to the robot (via the hidden operator) controlling the direction of the conversation. In the LSHG-S condition, Survivor Buddy displayed gaze behaviors, using the proposed method for the generation of head gaze based on sentence structure and time delays. The TSHG-S condition displayed gaze behaviors based on the semantic content of the dialog, which was similar to gaze behaviors exhibited in human–human conversation. In the NHG-S condition, Survivor Buddy looked directly at the participant throughout the interaction, without displaying any head gaze acts, and used only speech to interact.

B. Hypotheses

Three hypotheses were tested in this experiment.

1) *Hypothesis 1 (H1)*: Participants who interact with a robot exhibiting the LSHG-S condition will evaluate their experiences more positively than participants who interact with a robot exhibiting the NHG-S condition.
2) *Hypothesis 2 (H2)*: Participants who interact with a robot exhibiting the LSHG-S condition will evaluate the robot more positively than participants who interact with a robot exhibiting the NHG-S condition.
3) *Hypothesis 3 (H3)*: The LSHG-S condition's improvements over the NHG-S condition will be comparable with those of the TSHG-S condition.

C. Participants

The participants included 53 males and 40 females within an age range of 18–67 ($M = 32.38, SD = 14.84$). People with diverse backgrounds like students, engineers, administrators, firefighters, technicians, and doctors participated in the experiment. The ethnic backgrounds of the participants consisted of 68.8% Caucasian, 7.5% Asian, 14% Hispanic, 3.2% African American, and 6.5% Middle Eastern. Of the 93 participants, 64 reported owning a pet, and eight reported owning a robot. The participants' familiarity with robots was low ($M =$ 2.09, $SD =$ 1.67, $Md =$ 1, $IQR =$ 1 on a scale of 1 to 7). Video gaming experience (number of hours per week) was low ($M =$ 3.23, $SD =$ 5.7, $Md =$ 1, $IQR =$ 4).

D. Procedure

The experiment was conducted in three phases: 1) participant check-in, consent, and preinteraction questionnaires; 2) interaction with the robot; and 3) postinteraction questionnaires and debriefing. Phase 2 is described in detail below. (Phases 1 and 3 are straightforward.)

In phase 2, the participants were asked to lay in a simulated confined space: a wooden structure (0.91 m × 0.91 m × 2.44 m). The participant covered themselves with a sleeping bag, and the lights illuminating the pathway to the wooden structure in the experiment site were turned OFF. The participants saw a brief dramatic video, which showed a first-person view of a parking garage collapse. The video ended with flashes of light, followed by lighting sufficient to see the robot. At this point in the experiment, the simulation of a collapsed parking structure disaster environment and the robot system were made visible to the participant (see Fig. 3). The Survivor Buddy robot system was at a distance of 1.22 m, within the participant's personal zone [40], so as to increase the likelihood of social interaction. The screen displayed only a Survivor Buddy logo so that the only social cues were voice and head gaze acts. The interaction with the participant lasted approximately 15 min, and the robot followed a predefined script consisting of questions and simple directions. The robot could also repeat portions of the script upon request. The robot supervisor (hidden from view) would activate the text for the robot's turns in the dialog. The dialog initially followed the 911 dispatch protocol focusing on assessing the victim's physical health, gaining information about the location, and the nature of the event. Then, the dialog shifted to questions that assessed mental function. For example, "Paper is used commonly for writing. Can you name as many alternative uses as you can?" The interaction concluded with the participant being informed that rescuers had arrived, and Survivor Buddy's head closing to signify no further engagement.

E. Experimental Measures

The experiment used preinteraction and postinteraction questionnaires to evaluate the social acceptance of the robot. The preinteraction questionnaire consisted of nine attributes regarding the participant's age, occupation, gender, education level, prior robot experience, ethnicity, prior video gaming experience, robot ownership, and pet ownership. For this experiment,

TABLE IV
LIST OF ITEMS AND RELIABILITY FOR THE 23 MEASURES

Category	**Measure**	**Item**	**Cronbach's α**
Participants' Affective State	SAM: Valence [42]	"How positive/negative did you feel about your interaction with the robot presented?"	-
	Creativity [43]	Summation of the the total number of alternate uses participants generated within 30 s for three items: "shoe," "sheet of paper," and "license plate" during the interaction.	-
	Memory [44]	Summation of the the total number of memorized items recalled by the participant during the interaction. The robot read off 20 different memory items ("vacuum," "cat," "doorknob," etc.) then the robot diverted the participants attention for 30 s. The robot then allowed 30 s for participants to state as many of the items as they could remember.	-
	Person at Ease [45]	Index of four items: "I was scared," "I felt stressed," "I felt frustrated," and "I felt trapped."	.71
	SAM: Arousal [42]	"How agitated/comforted did you feel about your interaction with the robot presented?"	-
	Chance of Rescue [45]	Index of three items: "I was confident the rescuers would find me," "I believed that rescuers were on their way," and "I felt optimistic."	.78
Participants' Perception of the Robot	Robot Empathy [45]	Index of five items: "kind," "sincere," "empathetic," "sympathetic," and "concerned about me."	.76
	Robot Loyalty [45]	Index of four items: "the robot's primary purpose was to help me," "the robot would only do things that were in my best interest," "the robot was more loyal to me than the rescuers," and "the robot was on my side."	.76
	Robot Integrity [45]	Index of five items: "likeable," "trustworthy," "helpful," "honest," and "reliable."	.77
	Robot Caring [45]	Index of five items: "the robot liked me," "the robot saw the situation from my perspective," "the robot was concerned about me," "the robot was empathetic," and "the robot wanted me to be rescued."	.75
	Robot Engagement [6], [13], [46]	"The robot was engaging."	-
	Robot Likeability [45]	Index of five items: "I liked the robot," "the robot was friendly," "the robot made me feel relaxed," "I trusted the robot," and "the robot made me feel safe."	.87
	Human-Like Behavior [7]	"The robot behaved human-like."	-
	Robot Intelligence [7]	"Intelligent."	-
	Robot Detachment [45]	Index of three items: "humorless," "unemotional," and "cold."	.53 (unreliable)
	Robot Confidence [45]	Index of three items: "confident," "in control," and"masculine."	.26 (unreliable)
	Robot Competence [45]	Index of eight items: "committed to the task," "competent," "experienced," "informed," "intelligent," "qualified," "skilled," and"trained."	.84
	Robot Unpleasantness [45]	Index of seven items: "difficult to use," "dishonest," "incompetent," "rude," "unhelpful," "unkind," and "unpleasant."	.79
	Robot Extraversion [45]	Index of seven items: "outgoing," "extraverted," "vivacious," "jovial," "enthusiastic," "cheerful," and "perky."	.71
Consistency and Appropriateness of the Robot's Head Movements	Understandability of Robot Behaviors [46]	Index of three items: "I always knew what object the robot looked at," "I could easily tell which objects the robot looked at," and "I could understand the robot."	.86
	Gaze-Speech Synchronization [46]	"The robot synched its movements with what it was saying."	-
	Looking at Objects at Appropriate Times [46]	"The robot looked at the objects at appropriate times."	-
	Natural Movement [7], [46]	"The robot movements were natural."	-

The unreliable measures were not analyzed further.

23 measures were used to assess the interaction. The postinteraction questionnaire consisted of 21 measures modeled after [7], [40], [45], [46], and two objective measures, Memory and Creativity, captured during the interaction. Creativity and Memory were measured as part of this study because research has demonstrated that these measures inversely correlate with stress [47], [48]. Stress indicates arousal, which is a dimension of affect [49]. Two questions from the Self Assessment Manikin (SAM) [42] used a nine point Semantic Differential scale. The rest of the questions used a seven-point Likert scale, with one indicating strong disagreement or strongly negative, and seven indicating strong agreement or strongly positive. Measures with multiple items were checked for internal consistency using a Cronbach's Alpha statistic [50] (Cronbach's $\alpha > 0.70$ is considered to be reliable). Several questions were reverse coded to prevent participants from uniformly selecting a single rating. To better understand and interpret the results from such a large set of measures, the measures were categorized into one of three groups—*Participants' Positive Affective State, Participants' Perception of the Robot*, and the *Consistency and Appropriateness of the Robot's Head Movements* as shown in Table IV. Table IV reports the list of items and their reliability with regard to the 23 measures used to evaluate the performance of LSHG-S.

VI. DATA ANALYSIS AND RESULTS

This section presents the details of the data analysis, resulting statistics, and evaluations of the proposed hypotheses.

TABLE V
SUMMARY OF THE RESULTS

	Measure	Main Effect	Post-Hoc Results			Mean (*M*), Standard Deviation (*SD*), Median *(Md),* Interquartile Range *(IQR)*												Levene's Test
			LSHG-S vs NHG-S	TSHG-S vs NHG-S	LSHG-S vs TSHG-S	LSHG-S				TSHG-S				NHG-S				
Participants' Affective State	SAM: Arousal	$F[2, 90] = 8.43$, $p < .001$, $\eta^2 = .16$	$t(90) = 3.63$, $p = .001$, $d = .77$	$t(90) = 3.48$, $p = .002$, $d = .73$	$p = .99$	7.68	1.42	8.00	2.00	7.61	1.70	8.00	2.00	6.09	1.97	6.00	3.00	.15
Participants' Perception of the Robot	Robot Likeability	$F[2, 90] = 6.75$, $p = .002$, $\eta^2 = .13$	$t(90) = 3.05$, $p = .008$, $d = .64$	$t(90) = 3.33$, $p = .004$, $d = .70$	$p = .97$	4.93	.60	4.91	1.20	4.97	.63	5.00	1.40	4.36	.89	4.40	1.40	.10
	Human-Like Behavior	$F[2, 90] = 8.9$, $p < .001$, $\eta^2 = .17$	$t(90) = 4.03$, $p < .001$, $d = .85$	$t(90) = 3.10$, $p = .007$, $d = .65$	$p = .63$	5.10	1.42	5.00	2.00	4.74	1.34	5.00	2.00	3.54	1.74	3.00	3.00	.11
Consistency and Appropriateness	Understanding Robot Behavior	$F[2, 90] = 18.09$, $p < .001$, $\eta^2 = .56$	$t(90) = 4.63$, $p < .001$, $d = .98$	$t(90) = 5.29$, $p < .001$, $d = 1.12$	$p = .99$	5.57	.95	5.33	1.67	5.63	1.59	6.16	2.75	2.83	1.15	3.00	1.58	.22
	Gaze-Speech Synchronization	$F[2,90] = 47.87$, $p < .001$, $\eta^2 = .52$	$t(90) = 8.66$, $p < .001$, $d = 1.83$	$t(90) = 8.28$, $p < .001$, $d = 1.75$	$p = .93$	5.90	1.27	6.00	2.00	5.78	1.08	6.00	2.00	2.93	1.63	3.00	2.00	.24
	Looking at Objects at Appropriate Times	$F[2,90] = 14.6$, $p < .001$, $\eta^2 = .54$	$t(90) = 4.82$, $p < .001$, $d = 1.02$	$t(90) = 5.12$, $p < .001$, $d = 1.08$	$p = .98$	5.72	1.67	6.00	1.00	6.10	.99	6.00	1.25	3.28	1.43	3.00	2.25	.38
	Natural Movement	$F[2,90] = 16.69$, $p < .001$, $\eta^2 = .27$	$t(90) = 4.79$, $p < .001$, $d = 1.01$	$t(90) = 5.19$, $p < .001$, $d = 1.09$	$p = .92$	4.70	1.62	5.00	2.00	4.83	1.60	5.00	2.00	2.77	1.38	3.00	3.00	.57
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

TABLE VI
INTERPRETATION OF EFFECT SIZE

Effect Size	**Eta-Squared** (η^2)	**Cohen's** (d)
Small	0.01–0.06	0.20–0.49
Medium	0.06–0.14	0.50–0.79
Large	0.14+	0.80+

TABLE VII
RESULTS FOR HYPOTHESES H1, H2, AND H3 BY CATEGORY

Category	**Hypothesis**	**Results for Bonferroni Correction**
Participants' Affective State	H1	Support for SAM: Arousal (LSHG-S versus NHG-S—$p = 0.001$).
	H3	Both LSHG-S and TSHG-S showed significant improvements over NHG-S in SAM: Arousal (LSHG-S versus NHG-S—$p = 0.001$, TSHG-S versus NHG-S—$p = 0.002$).
Participants' Perception of the Robot	H2	Support for Robot Likeability (LSHG-S versus NHG-S—$p = 0.008$) and Human-Like Behavior (LSHG-S versus NHG-S—$p < 0.001$).
	H3	Both LSHG-S and TSHG-S showed significant improvements over NHG-S in Robot Likeability (LSHG-S versus NHG-S—$p = 0.008$, TSHG-S versus NHG-S—$p = 0.004$) and Human-Like Behavior (LSHG-S versus NHG-S—$p < 0.001$, TSHG-S versus NHG-S—$p = 0.007$).
Consistency and Appropriateness of the Robot's Head Movements	H2	Support for Understanding Robot Behavior (LSHG-S versus NHG-S—$p < 0.001$), Gaze-Speech Synchronization (LSHG-S versus NHG-S—$p < 0.001$), Looking at Objects at Appropriate Times (LSHG-S versus NHG-S—$p < 0.001$), and Natural Movement (LSHG-S versus NHG-S—$p < 0.001$).
	H3	Both LSHG-S and TSHG-S showed significant improvements over NHG-S in Understanding Robot Behavior (LSHG-S versus NHG-S—$p < 0.001$, TSHG-S versus NHG-S—$p < 0.001$), Gaze-Speech Synchronization (LSHG-S versus NHG-S—$p < 0.001$, TSHG-S versus NHG-S—$p < 0.001$), Looking at Objects at Appropriate Times (LSHG-S versus NHG-S—$p < 0.001$, TSHG-S versus NHG-S—$p < 0.001$), and Natural Movement (LSHG-S versus NHG-S—$p < 0.001$, TSHG-S versus NHG-S—$p < 0.001$).

A. Data Analysis

The data analysis consisted of an univariate ANOVA for each reliable measure and a posthoc analysis using Tukey's HSD. There was homogeneity of variances for each of the measures, as assessed by Levene's Test of Homogeneity of Variance. The assumption of normality is met because the Central Limit Theorem [51] states that the sample means will be normally distributed for sample sizes greater than 30 (the sample size for the study was 93), even if the population is positively skewed, negatively skewed, or even binomial. In order to avoid Type I error, the Bonferroni correction [52] for multiple testing was applied. The corrected significance level after the Bonferroni correction [52] is $p < 0.0024$ $(0.05/21)$.

An ANCOVA was performed to further account for other potential sources of variance like Gender, Age, Video Gaming Experience, Robot OwnerShip, Pet Ownership, Past Experience with Robots, and Ethnicity. Table V lists the main effects (column 2) from the ANOVA, Tukeys' posthoc results between the conditions (columns 3–5) for the seven significant measures (column 1), the mean, standard deviation, median, and interquartile range for each of the measures (columns 6–17), and the Levene's Test of Homogeneity of Variance (column 18). The eta-squared (η^2) and Cohen's d effect sizes can be interpreted using the scale [53] shown in Table VI. Table VII summarizes the results for each hypothesis.

B. Results for Social Acceptance

The three categories of measures, i.e., *Participants' Positive Affective State, Participants' Perception of the Robot,* and *Consistency and Appropriateness of the Robot's Head Movements*,

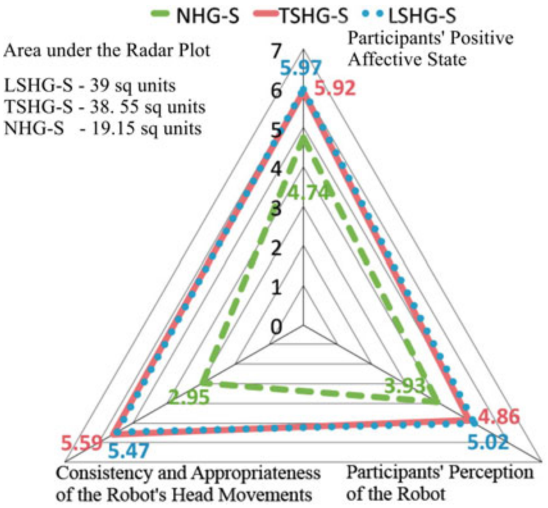

Fig. 4. Radar plot comparing the means of TSHG-S, LSHG-S, and NHG-S conditions using Bonferroni Correction. The area under the graph is Social Acceptance.

when viewed holistically, give insight into the social acceptance of each head gaze generation method. To develop an intuitive understanding of the results, the means for each condition of the three categories of measures were represented on a radar plot. Fig. 4 illustrates the results with Bonferroni correction. The area under the radar plot (see Fig. 4) reflects a measure of overall social acceptance. Both LSHG-S and TSHG-S conditions have comparable areas with similar shapes suggesting that the LSHG-S condition engendered high levels of social acceptance similar to the TSHG-S condition. The TSHG-S and LSHG-S conditions have a larger area under the graph when compared with the NHG-S condition, suggesting that they achieved greater social acceptance than NHG-S.

C. Influence of Covariates on the Measured Attributes

Using an ANCOVA, the influence of Gender, Age, Video Gaming Experience, Robot Ownership, Pet Ownership, Past Experience with Robots, and Ethnicity on the 21 measured attributes was analyzed.

1) Participants who had more experience with robots got more agitated with the robot ($F[1,84] = 4.82$, $p = 0.03$, $\eta^2 = 0.05$) and found the robot to be less Extraverted ($F[1,84] = 9.98$, $p < 0.001$, $\eta^2 = 0.11$).
2) Increased age led to a higher perception of a better Chance of Rescue ($F[1,84] = 8.97$, $p = 0.003$, $\eta^2 = 0.09$) by a robot.
3) Males rated the robot as being more unpleasant than females ($F[1,84] = 5.42$, $p = 0.02$, $\eta^2 = 0.06$).
4) Caucasians viewed the robot to be more extraverted than Asians ($F[1,84] = 2.58$, $p = 0.04$, $\eta^2 = 0.10$).
5) Increased video gaming experience resulted in increased perception of Natural Movement ($F[1,84] = 11.56$, $p = 0.03$, $\eta^2 = 0.04$) and Robot Synchronization ($F[1,84] = 11.02$, $p = 0.001$, $\eta^2 = 0.06$).

VII. Discussion

The results show that social robots can achieve adequate and appropriate head gaze-speech synchronization by exploiting affordances in the sentence structure of dialog and time delays.

A. Interpretation and Discussion of the Results

The results from this experiment revealed that LSHG-S was socially acceptable for human–robot interaction. LSHG-S performs at least, as well as the TSHG-S, when compared with NHG-S and is preferred to NHG-S for the following seven attributes: SAM: Arousal, Robot Likeability, Human-Like Behavior, Understanding Robot Behavior, Gaze-Speech Synchronization, Looking at Objects at Appropriate Times, and Natural Movement. The results for *Consistency and Appropriateness of the Robot's Head Movements* are particularly strong and suggest that head gaze behaviors generated by the robot were consistent, appropriate, and sensitive to the context. The results reinforce findings from Kirchhof & Ruiter [36] that gesture comprehension is temporally more flexible than gesture production, and participants tolerate loose synchronization of head gaze with speech.

These results are surprising because LSHG-S did not use semantic understanding to precisely match head gaze with dialog. The findings from this experiment suggests that 1) affordances in the sentence structure of dialog and time delays that were developed as a part of this research effort are adequate and acceptable for human–robot interaction, and 2) the LSHG-S is a more viable and practical method for the generation of head gaze than the commonly used TSHG-S.

While the overall results are significant for social robotics in general, it has strong applicability to eldercare and therapeutic robotics, in addition to the victim management application presented in this paper. A rescue operation might take up to 6–10 h after a victim is located [39], during which it is psychologically helpful for the "victim" to perceive the robot to be social and following human interpersonal communication norms [40].

B. Four Factors Relevant for User Perception of Speech-Gaze Synchronization

This paper identifies four factors that are important for user perception of speech-gaze synchronization, which could also explain why the LSHG-S condition performed well.

1) *Gesture comprehension is temporally more flexible than gesture production*: The gesture-speech synchrony might be a consequence of the production system, but may not be essential for comprehension [36]. If people were very sensitive to TSHG-S, then the results would have indicated that the participants would have a preference for the TSHG-S over LSHG-S, which was not the case. The semantic and temporal flexibility of head gaze production with robots requires further investigation.
2) *Expectation of gaze in a semihumanoid robot*: The expectations of gaze from a robot with a low degree of anthropomorphism and a more mechanical appearance may not be as high as those required for human interpersonal

interactions. Survivor Buddy has a low degree of anthropomorphism, with a mechanical appearance. It does not have eyes, hands, or even the shape of a human head. We believe that even though Survivor Buddy has an intelligent conversation with the participants, they do not hold it to the same high standards required in human interpersonal communication, or even human-android communication.

3) *Importance of synchronization at the start and end of turns is greater than at the middle of turns*—Yamazaki *et al.* [9] discuss that timing is critical at the start and end of turns for conversation. Cassell *et al.* [18] emphasize the importance of middle of turn events for superior performance; however, the timing of these events has not been examined. The proposed method has precise synchronization at the start and end of turns; however, it approximates the semantic structure during the middle of turns to support autonomous generation. These approximations did not adversely impact the participants, and resulted in performance similar to that of TSHG-S. The timing of the middle of turn events may be of relatively low importance.
4) *Absence of lips*: Humans perceive speech-lip asynchrony as unnatural [54], and prior work suggests that speech needs to be tightly synchronized with lips, but not with gestures [36]. In the experiment, the Survivor Buddy robot did not have lips. This could possible explain why the participants did not perceive LSHG-S to be unnatural, but instead found it to be equivalent to that of TSHG-S.

C. Limitations of the Experiment

The current implementation has four limitations, which present opportunities for future work.

1) *Implementation used a priori sensor data*: The robot did not have object recognition or speech recognition capabilities. This is because this research focused on the generation of appropriate head gaze behaviors, not recognition of objects or speech, which are challenging research areas in their own right. However, these limitations can be expected to be addressed in the future.
2) *Limited to head gaze, no eye gaze*: The implementation generated only head gaze acts and ignored eye gaze. However, the inference from the sentence structure and timing could readily extend to eye gaze, if the robot had mechanical or virtual eyes. Of the two body components, head gaze may be more immediately valuable as many robots do not have eyes capable of gaze. For example, robot animals often have fixed cameras or lights for eyes, and nonanthropomorphic robots do not have eyes at all.
3) *Single domain validation*: The experiment showed that gaze primitives validated in other domains (TSHG-S condition) had positive results for victim management, as shown by the high rating of the TSHG-S condition. Because the performance of the LSHG-S condition (which used approximations) was as good as the TSHG-S condition for victim management in a search and rescue domain, the expectation is that the results will transfer to other domains. The next step is to validate the system in the field using a very realistic setting. Validation of the system in other domains can be explored in future work.
4) *Controlled conversational interaction*: The conversational interaction was directed by the robot (using a hidden operator). This method was adopted because speech recognition tools are unreliable and affect the repeatability and consistency of the experiment across conditions. Future speech recognition systems will permit the participants to have a collaborative dynamic conversation with the robot.

VIII. Conclusion

This paper is the first to propose and evaluate the effects of LSHG-S for five functions: *communicating attention, engaging in a verbal conversation, exhibiting aliveness, showing various mental states*, and *referential gaze* to objects in the environment. LSHG-S was realized using eight novel affordances for turn taking and semantics from the sentence structure, time delays, and nine production rules. The proposed method was implemented on an affective robot for victim management and evaluated in a user study. The results from the experiment demonstrated high efficacy of LSHG-S in the key attributes of SAM: Arousal, Robot Likeability, Human-Like Behavior, Understanding Robot Behavior, Gaze-Speech Synchronization, Looking at Objects at Appropriate Times, and Natural Movement. Additionally, LSHG-S performs at least, as well as the TSHG-S, when compared with NHG-S for each of these seven measures. The overall results indicate that the affordances developed as a part of this research effort are adequate and socially acceptable for human–robot interaction.

The proposed method contributes three benefits for social robotics. First, it supports autonomous annotation of head gaze in text, that is independent of the semantics of the dialog. The method reduces the workload of researchers, as they are no longer required to tediously hand code every scenario, Wizard-of-Oz style. Second, the robot can generate socially acceptable head gaze behaviors in real time for very open-ended interactive scenarios. These advantages are very important in situations where robot responses cannot be anticipated *a priori* (e.g., personal robots for eldercare). Third, the synchronization of head gaze with speech is flexible and promotes practical implementation in robots with lesser capabilities. In short, LSHG-S is efficient, effective, and flexible.

This research also contributes to the general understanding of human–robot interaction, in particular, addressing the question—*where is less competence tolerable?* as well as the practical implementation of the theory. The experiment raises two new research questions: *What is the extent to which humans tolerate loose synchronization of head gaze and speech?* and *What factors affect expected synchronization of head gaze-speech?* For example, What role does people's (perceived) appearance of the robot have on head gaze-speech synchronization? Additional studies need to be conducted to address these questions.

ACKNOWLEDGMENT

The authors would like to thank Z. Henkel and J. Suarez for their helpful comments and suggestions.

REFERENCES

[1] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase, "Robot mediated round table: Analysis of the effect of robot's gaze," in *Proc. 11th IEEE Int. Workshop Robot Human Interact. Commun.*, 2002, pp. 411–416.

[2] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Proc. Int. Conf. Intell. Robots Syst.*, 2003, pp. 708–713.

[3] T. Minato, M. Shimada, H. Ishiguro, and S. Itakura, "Development of an android robot for studying human-robot interaction," *Innov. Appl. Artif. Intell.*, vol. 3029, pp. 424–434, 2004.

[4] D. Sakamoto, T. Kanda, T. Ono, M. Kamashima, M. Imai, and H. Ishiguro, "Cooperative embodied communication emerged by interactive humanoid robots," in *Proc. 13th IEEE Int. Workshop Robot Human Interact. Commun., RO-MAN*, Sep 2004, pp. 443–448.

[5] K. MacDorman, T. Minato, M. Shimada, S. Itakura, S. Cowley, and H. Ishiguro, "Assessing human likeness by eye contact in an android testbed," in *Proc. 27th Annu. Meet. Cognitive Sci. Soc.*, 2005, pp. 21–23.

[6] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artif. Intell.*, vol. 166, pp. 140–164, 2005.

[7] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *Proc. IEEE Int. Conf. Humanoid Robots*, 2006, pp. 518–523.

[8] Y. Kuno, K. Sadazuka, M. Kawashima, K. Yamazaki, A. Yamazaki, and H. Kuzuoka, "Museum guide robot based on sociological interaction analysis," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 1191–1194.

[9] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka, "Precision timing in human-robot interaction: coordination of head movement and utterance," in *Proc. 26 Annu. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 131–140.

[10] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Adapting robot behavior for human–robot interaction," *IEEE Trans. Robot.*, vol. 24, no. 4, pp. 911 –916, Aug. 2008.

[11] B. Mutlu, T. Shiwa, T. K, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," *Proc. 4th ACM/IEEE Int. Conf. Human Robot Interact.*, 2009, pp. 61–68.

[12] M. Shimada, Y. Yoshikawa, M. Asada, N. Saiwaki, and H. Ishiguro, "Effects of observing eye contact between a robot and another person," *Int. J. Social Robot.*, vol. 3, pp. 143–154, 2011.

[13] A. Holroyd, C. Rich, C. L. Sidner, and B. Ponsler, "Generating connection events for human-robot collaboration," in *Proc. RO-MAN IEEE*, Jul. 31–Aug. 3, 2011, pp. 241–246.

[14] C. Huang and B. Mutlu, "Robot behavior toolkit: Generating effective social behaviors for robots," in *Proc. 7th ACM/IEEE Int. Conf. Human-Robot Interact.*, 2012, pp. 25–32.

[15] C.-M. Huang and B. Mutlu, "The repertoire of robot behavior: Designing social behaviors to support human-robot joint activity," *J. Human-Robot Interact.*, vol. 2, no. 2, pp. 80–102, 2013.

[16] Y. Matsusaka, S. Fujie, and T. Kobayashi, "Modeling of conversational strategy for the robot participating in the group conversation," in *Annu. Conf. Int. Speech Commun. Assoc.*, 2001, pp. 2173–2176.

[17] C. T. Ishi, C. Liu, H. Ishiguro, and N. Hagita, "Head motions during dialogue speech and nod timing control in humanoid robots," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, 2010, pp. 293–300.

[18] J. Cassell, O. Torres, and S. Prevost, "Turn taking vs. discourse structure: How best to model multimodal conversation," in *Machine Conversations*. Norwell, MA, USA: Kluwer, 1998.

[19] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proc. 4th ACM/IEEE Int. Conf. Human Robot Interaction.*, 2009, pp. 69–76.

[20] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," in *Proc. 10th IEEE Int. Workshop Robot Human Interact. Commun.*, 2001, pp. 512 –517.

[21] M. Staudte and M. Crocker, "The effect of robot gaze on processing robot utterances," in *Proc. 31th Annu. Conf. Cognitive Sci. Soc.*, Amsterdam, The Netherlands, 2009, pp. 431–436.

[22] M. Staudte, and M. W. Crocker, "Visual attention in spoken human-robot interaction," in *Proc. 4th ACM/IEEE Int. Conf. Human Robot Interact.*, 2009, pp. 77–84.

[23] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *J. Pers. Social Psychol.*, vol. 23, no. 2, pp. 283–292, 1972.

[24] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[25] A. S. Meyer, A. M. Sleiderink, and W. J. Levelt, "Viewing and naming objects: Eye movements during noun phrase production," *Cognition*, vol. 66, no. 2, pp. B25–B33, 1998.

[26] Z. M. Griffin, "Gaze durations during speech reflect word selection and phonological encoding," *Cognition*, vol. 82, no. 1, pp. B1–B14, 2001.

[27] R. R. Murphy, *Introduction to AI Robotics*. Cambridge, MA, USA: MIT Press, 2000.

[28] M. Cherubini, R. de Oliveira, N. Oliver, and C. Ferran, "Gaze and gestures in telepresence: Multimodality, embodiment, and roles of collaboration," in *CoRR*, 2010.

[29] C. Liu, C. T. Ishi, H. Ishiguro, and N. Hagita, "Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction," in *Proc. 7th Annu. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2012, pp. 285–292.

[30] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies: Why and how," in *Proc. 1st Int. Conf. Intell. User Interfaces.*, 1993, pp. 193–200.

[31] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge, U.K.: Cambridge Univ. Press, 1976.

[32] E. Goffman, "Footing," *Semiotica*, vol. 25, nos. 1/2, pp. 1–30, 1979.

[33] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[34] V. Srinivasan, C. Bethel, R. Murphy, and C. Nass, "Validation of a behavioral robotics framework for head social gaze," in *Proc. Workshop Gaze Human Robot Interact., From Model. Commun.*, 2012.

[35] M. Halliday, *Intonation and Grammar in British English* (ser. Janua linguarum: Series practica). Berlin, Germany: Mouton, 1967.

[36] C. Kirchhof and J. P. D. Ruiter, "On the audiovisual integration of speech and gesture," presented at the Proc. 5th Conf. Int. Soc. Gesture Studies, Lund, Switzerland, 2012.

[37] R. Murphy, A. Rice, N. Rashidi, Z. Henkel, and V. Srinivasan, "A multidisciplinary design process for affective robots: Case study of survivor buddy 2.0," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 701–706.

[38] V. Srinivasan and R. Murphy, "A survey of social gaze," in *Proc. 6th Int. Conf. Human-Robot Interact.*, 2011, pp. 253–254.

[39] T. Fincannon, L. Barnes, R. Murphy, and D. Riddle, "Evidence of the need for social intelligence in rescue robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2004, vol. 2, pp. 1089–1095.

[40] C. Bethel and R. Murphy, "Non-facial and non-verbal affective expression for appearance-constrained robots used in victim management," *Paladyn. J. Behav. Robot.*, vol. 1, pp. 219–230, 2010.

[41] C. L. Bethel and R. R. Murphy, "Non-facial/non-verbal methods of affective expression as applied to robot-assisted victim assessment," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2007, pp. 287–294.

[42] M. Bradley, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[43] J. P. Guilford, "Creativity: Its measurement and development," *A Source Book for Creative Thinking*. New York, NY, USA: Charles, 1962, pp. 151–167.

[44] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology*. New York, NY, USA: Teachers Coll., Columbia Univ., 1913, no. 3.

[45] V. Groom, V. Srinivasan, C. Bethel, R. Murphy, L. Dole, and C. Nass, "Responses to robot social roles and social role framing," in *Proc. Int. Conf. Collab. Technol. Syst.*, May 2011, pp. 194–203.

[46] C. Rich, B. Ponsleur, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, 2010, pp. 375–382.

[47] W. H. Teichner, E. Arees, and R. Reilly, "Noise and human performance, a psychophysiological approach," *Ergonomics*, vol. 6, no. 1, pp. 83–97, 1963.

[48] L. Schwabe and O. T. Wolf, "Learning under stress impairs memory formation," *Neurobiol. Learn. Memory*, vol. 93, no. 2, pp. 183–188, 2010.

[49] J. A. Russell, "Evidence of convergent validity on the dimensions of affect," *J. Pers. Social Psychol.*, vol. 36, no. 10, pp. 1152–1168, 1978.

[50] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[51] J. Rice, *Mathematical Statistics and Data Analysis*. Boston, MA, USA: Cengage Learning, 2006.

[52] J. P. Shaffer, "Multiple hypothesis testing," *Annu. Rev. Psychol.*, vol. 46, no. 1, pp. 561–584, 1995.

[53] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Evanston, IL, USA: Routledge, 1988.

[54] A. Vatakis, J. Navarra, S. Soto-Faraco, and C. Spence, "Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments," *Exp. Brain Res.*, vol. 185, no. 3, pp. 521–529, 2008.

Vasant Srinivasan is currently working toward the Ph.D. degree with the Computer Science and Engineering Department, Texas A&M University, College Station, TX, USA.

Cindy L. Bethel received the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, FL, USA, in 2009.

She is an Assistant Professor with the Computer Science and Engineering Department, Mississippi State University (MSU), Mississipi State, MS, USA. She is the Director of the Social, Therapeutic, and Robotic Systems (STaRS) Lab and a Research Fellow with the MSU Center for Advanced Vehicular Systems Human Performance Group. She was a NSF/CRA/CCC Computing Innovation Postdoctoral Fellow in the Social Robotics Laboratory, Yale University. Her research interests include human–robot interaction, affective computing, robotics, human–computer interaction and interface design, artificial intelligence, and psychology. Her research focuses on applications associated with the use of robots for therapeutic support, law enforcement, search and rescue, and military.

Robin R. Murphy (F'10) received the B.M.E. degree in mechanical engineering, the M.S. and Ph.D. degrees in computer science in 1980, 1989, and 1992, respectively, from Georgia Tech, Atlanta, GA, USA.

She was a Rockwell International Fellow. She is the Raytheon Professor of Computer Science and Engineering with Texas A&M, College Station, TX, USA, the Director of the Center for Robot-Assisted Search and Rescue, and the Center for Emergency Informatics. She has more than 150 publications on artificial intelligence, human–robot interaction, and robotics including two textbooks, *Introduction to AI Robotics* and *Disaster Robotics*.