How Was Your Day? Evaluating a Conversational Companion

David Benyon, Björn Gambäck, Preben Hansen, Oli Mival, and Nick Webb

Abstract—The "How Was Your Day" (HWYD) companion is an embodied conversational agent that can discuss work-related issues, entering free-form dialogues while discussing issues surrounding a typical work day. The open-ended nature of these interactions requires new models of evaluation. Here, we describe a paradigm and methodology for evaluating the main aspects of such functionality in conjunction with overall system behavior, with respect to three parameters: functional ability (i.e., does it do the "right" thing conversationally), content (i.e., does it respond appropriately to the semantic context), and emotional behavior (i.e., given the emotional input from the user, does it respond in an emotionally appropriate way). We demonstrate the functionality of our evaluation paradigm as a method for both grading current system performance, and targeting areas for particular performance review. We show correlation between, for example, automatic speech recognition performance and overall system performance (as is expected in systems of this type), but beyond this, we show where individual utterances or responses, indicated as positive or negative, characterize system performance, and demonstrate how our combination evaluation approach highlights issues (both positive and negative) in the companion system's interaction behavior.

Index Terms—Companions, embodied conversational agents, evaluation, appropriateness of dialogue

1 Introduction

Companion technologies, or companions, refer to an emerging form of interaction between people and technologies [1]. The term "companion" has been used naturally, without specific definition, in the EU projects Semaine (www.semaine-project.eu) and LIREC (www.lirec.eu), which explored the design of digital and robotic companions. Wilks [2] has characterized the companion concept as a personalized conversational, multimodal interface, one that knows its owner and is implemented on a range of platforms, both static and mobile. For us, companions are advanced spoken language dialogue systems that attempt to go beyond the limited functionality of current task-oriented dialogue systems by being cooperative, collaborative dialogue partners and forming long-term relationships with the people who interact with them. Uniting these views is that companions foreground the social and affective, unlike other approaches to interaction that foreground the efficiency, effectiveness, or utility of technologies.

- *D. Benyon and O. Mival are with the Centre for Interaction Design, Edinburgh Napier University, 10 Colinton Road, Edinburgh, Scotland EH10 5DT, United Kingdom. E-mail: {d.benyon,o.mival}@napier.ac.uk.*
- *B. Gambäck is with the Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim 7491, Norway. E-mail: gamback@idi.ntnu.no.*
- *P. Hansen is with the Department of Computer and Systems Sciences, Stockholm University, Kista SE-164 29, Sweden. E-mail: preben@dsv.su.se.*
- *N. Webb is with the Department of Computer Science, Union College, 807 Union Street, Schenectady, NY 12308. E-mail: webbn@union.edu.*

Manuscript received 23 Apr. 2012; revised 28 May 2013; accepted 5 June 2013; published online 13 June 2013.
Recommended for acceptance by C. Pelachaud.
For information on obtaining reprints of this article, please send E-mail to: taffc@computer.org and reference IEEECS Log Number TAFFC-2012-04-0031.
Digital Object Identifier no. 10.1109/T-AFFC.2013.15.

The volume edited by Wilks [1] contains 24 chapters devoted to wide-ranging discussions of companion issues, from falling in love with a companion to being a Victorian companion. Benyon and Mival [3] argue that companions aim to change human-computer interaction (HCI) into human-companion relationships. This builds upon the ideas of affective computing [4] and designing for the emotional involvement of people with technologies. Companions may be represented as "virtual human" on-screen characters or as embodied conversational agents (ECA), but do not have to be. Companions encompass the widest possible range of devices and forms of interaction that woven together produce a relationship-building experience for people.

Companions represent the maturation of a concept that has been in computing and HCI since it earliest days. Weizenbaum [5] developed his program ELIZA in the late 1960s. It was able to hold remarkably convincing conversations (through a type-written interface) that many people clearly found engaging. Chatbots are available online that perform a similar function. However, these systems do not have any underlying purpose other than to chat. Companions and ECAs are interactive technologies that aim to achieve higher conversational or relationship goals. They draw upon research into software engineering, speech and language, artificial intelligence, and theories of conversation and interaction, politeness, humor, personality, emotion, trust, social presence as well as other theories of human relations [6]. Companions are in particular designed to develop and maintain longer term socio-emotional relations with users or to focus the establishment of trust, for example, by utilizing conversational strategies to support humor [7], politeness [8], or empathy [9].

The challenge that we address in this paper is how should companions be evaluated. We present the detailed

Fig. 1. The "How Was Your Day" companion interface.

evaluation of a specific instantiation of the companion concept. This is the "How Was Your Day?" (HWYD) companion, an ECA developed to have engaging conversational interactions with the typical office worker returning home after work. The HWYD system was developed as part of the EU funded integrated project COMPANIONS investigating the companion concept. Several other companion concepts were developed including a senior companion aimed at having conversations with older people and a health and fitness companion aimed at encouraging people to exercise more. The aim of the HWYD companion was to showcase a sophisticated model of natural dialogue interaction using a speech-based interface. The scenario was chosen because it was seen to be representative of the sort of deployment that companions in the home could offer in a few years' time. A large screen in the home would display the avatar who would greet the users when they got home, and be capable of holding conversations around 40 office-related topics in the context of asking about a person's working day [10].

The aim of the laboratory-based evaluation of the companion was to provide summative data about the performance of the system across a range of representative scenarios and to explore the more general issues of evaluating companions. The system was evaluated in the form that the developers delivered it (Fig. 1) with the exception of one scenario (see Section 3). Although it is unusual for users to be able to see the workings of an ECA, the interface served to highlight the performance of the system in terms of accessed modules and of the recognized speech and emotions.[1]

The paper is laid out as follows. This section describes the HWYD companion system and discusses some previous efforts to evaluate spoken dialogue systems. Section 2 introduces the proposed evaluation paradigm for companions with its subjective and objective measures. Section 3 elaborates on the evaluation methodology and on how user studies were set up and performed. The scenarios adopted for those studies play a vital role in the evaluations and are described in detail in Section 4. Results of experimental user studies carried out along these lines are presented and analyzed in Section 5. Section 6 finally discusses the experiences from the experimental evaluations.

1.1 The "How Was Your Day" Companion

The user interface (UI) of the HWYD system [11] is illustrated in Fig. 1. On the left we see an avatar exhibiting facial expressions and gestures. The system is rendered on a HD screen with a roughly life-size ECA. The HWYD companion can engage in long, free-form conversations about events that have taken place during the user's working day. Part of such a conversation between the user and system can be seen in the middle of the figure. The right part of the figure shows the system architecture; the modules currently active during processing light up on the display.

The system both allows for user initiative and displays system initiative, including questions, comments, advice, and overall attempts to positively influence the user's emotional state. The user's emotional state is monitored through acoustic and linguistic information, allowing the system to generate affective spoken responses. Two distinct processing loops are invoked in order to keep the dialogue flow fast and natural. A "short" loop takes care of back-channel interaction in more or less real-time (<500 ms), allowing the companion to react to the emotional state of the user through facial expression, gestures, and short statements. More traditional dialogue management guides the "long" loop that gathers event representations from user statements and uses this to generate answers giving advice and providing comfort, typically in the form of a short tirade (4-5 utterances) from the companion.

The system processed the modules shown on the right-hand side of Fig. 1 in the following order. Nuance's dragon naturally speaking automatic speech recognition (ASR) sends the recognized words to dialogue act tagging, which along with acoustic analysis information and acoustic turn

1. A demonstration video of the HWYD companion in action is available at http://tinyurl.com/HWYD-companion.

taking (ATT), identifies the dialogue acts that are passed to natural language understanding (NLU). An emotional speech recognizer, EmoVoice (EV) [12] returns information indicating the arousal and valence of the acoustic properties of the user speech as negative-passive, negative-active, positive-active, positive-passive or neutral, while a text-based sentiment analyzer (SA) [13] operates on the ASR utterance transcript, classifying clauses as negative, neutral, or positive. The two emotional inputs are fused together by emotion modeling (EM): essentially, the speech-based valence category is overridden by sentiment analysis if EmoVoice's confidence score is below a preset threshold.

In the "long" loop, the rule-based dialogue manager (DM) takes the affect-annotated semantic output of the NLU and determines the next system turn, which is generated by the plan-based affective strategy module (ASM) [10] and handed to natural language generator (NLG). The NLG output is passed both to text-to-speech synthesis (an extension of the Loquendo TTS system including paralinguistic elements such as exclamations and laughter, and emotional prosody generation for negative and positive utterances), and to the ECA module guiding the movements of the avatar, producing gestures, and facial expressions conveying the companion's emotional state.

1.2 Evaluating Companions

Companions are targeted as persistent, collaborative, conversational partners, where the user may have a wide degree of initiative in the resulting interaction. Rather than singular, focused tasks, as seen in the majority of deployed dialogue systems, fully developed companions can have a range of tasks and be expected to switch task on demand. In the case of the HWYD companion, the aim was to maintain the dialogue across a range of conversational topics. When devising an evaluation paradigm for such systems, the successful completion of dialogue tasks needs to be balanced with measures of both "conversational performance" and overall user experience. The assumption in traditional dialogue evaluation is that the quality of the conversation correlates with *user satisfaction*; if the resulting dialogue is annoying or repetitive, a corresponding drop in user satisfaction is expected. However, for companions, it is *overall user experience* that is important, covering the entire interaction. Poor text-to-speech performance or poor response time may have a disproportional effect on user experience. The notion of user experience goes beyond satisfaction to include both the pragmatic characteristics and the hedonic characteristics of interaction [14].

A significant amount of effort has been spent on evaluating spoken language dialogue systems, mostly relying on a combination of observable metrics and user feedback (cf. [15], [16], [17]). Efficiency and effectiveness metrics often include the number of user turns, system turns, and total elapsed time. For the "quality of interaction," it is usual to record speech recognition rejections, time out prompts, help requests, barge-ins, mean recognition score (concept accuracy), and cancellation requests, all being somewhat functional descriptors of the quality of interaction.

The DARPA communicator program made extensive use of the PARADISE (PARAdigm for dialogue system evaluation) metric [18], which was developed to evaluate the performance of spoken dialogue systems, in a way decoupled from the task the system was attempting. "Performance" of a dialogue system is affected both by *what* the user and the dialogue agent working together accomplish, and by *how* it gets accomplished. PARADISE aims to maximize task completion, while simultaneously minimizing dialogue costs, evaluated as both objective efficiency of the dialogue (e.g., number of turns) and some qualitative measure. A consequence of this model is that often the dialogue quality parameters are tuned to overcome the deficiencies highlighted by the observable metrics, such as discussed by Hajdinjak and Mihelič [19]. For example, using explicit confirmation increases the likelihood of successful task completion in environments with low speech recognition performance, and so is often chosen, despite being regarded as a bit unnatural in comparative human-human speech data [20].

The lack of a community-wide method for evaluating conversational performance of spoken language dialogue systems acts as a barrier to the wholesale development of usable, practical systems beyond simple, task-oriented interaction. We want to explore a method of scoring conversational performance directly; measuring the system's capability to maintain a conversation based on the progression of the dialogue. We believe that conversational performance can be measured in terms of appropriateness, and indeed several researchers previously looked at using a mechanism of appropriateness of dialogue as a measure of effective communication strategies (see [21], [22], [23], [24]).

2 Evaluation Paradigm

In order to evaluate a companion, some overall system properties need to be charted: functional ability (does it do the "right" thing?), content (does it respond appropriately to the semantic context?), and emotional behavior (given the emotional input from the user, does it respond in an emotionally appropriate way?). To this end, we have developed an evaluation process that considers, and correlates, three types of features [2]:

1. *Metric-centric.* The use of quantitative methods to determine values for dialogue metric data including word error rate of speech recognition and concept error rate of NLU, in conjunction with readily computable scores such as dialogue duration; number of turns; words per turn, and so on
2. *User-centric.* Qualitative methods used to acquire subjective impressions and opinions from the users of the companions prototypes, including Likert-based surveys, focus groups, and interviews.
3. *Measure of appropriateness.* An annotation of the data resulting from the metric-centric evaluation. Human labelers assign categories to both system and user utterances, with particular focus on system behavior. Labels capture the appropriateness of an utterance in the context of the on-going dialogue. For example, if the system asks a particular question, it may be

TABLE 1
Objective Dialogue Metrics

Dialogue Metrics	Dimensions	
Average utterance length (seconds)	user	system
Average delay (seconds)	user	system
Average turn duration (seconds)	user	system
Average words per turn	user	system
Total number of turns	user	system
Average number of user words	ASR	transcript
Overall error rate	word	concept
Total dialogue duration	seconds	utterances

judged to be appropriate, but if the system subsequently repeats the same question, when the user has provided a valid answer, the same utterance could be judged to be inappropriate in that context.

2.1 Objective Speech and Dialogue Metrics

For the objectively measurable quantities, a set of requirements must be met. In particular, standard timing information needs to be collected from each interaction. Delay times between utterances, both system and user, should be captured, as well as dialogue length, in time and in number of utterances. Word error rate (WER) and concept error rate (CER) are calculated based on deletions, insertions, and substitutions. The 16 objective metrics used in this study are outlined in Table 1.

2.2 Subjective Measures

Traditional dialogue systems place a high reliance on user feedback. Measures of how people relate to companions were collected through online questionnaires. The questions are organized around six themes developed following several empirical investigations of companion technologies: naturalness of the companion, utility of the companion, participant-companion relationship nature, emotion demonstrated by the companion, personality of the companion, and social attitudes of the companion. Some of these themes are geared toward specific behaviors of the companion system, for example, targeted questions on the use of emotion by the companion (both recognizing emotion from the user and generating appropriate emotion). The themes all contribute to people developing a sense of social presence of technologies, and encourage them to move from simply interacting with a system to forming a relationship with it, which is central to the notion of companions [3]. These themes, in conjunction with the objective metrics, allow us to assess the behavior of the companion as a conversational agent.

2.3 Measure of Appropriateness

Appropriateness is a measure of each utterance in a dialogue, where human annotators score each utterance, grading the level of information presented and the progression of the dialogue as appropriate or otherwise. The concept is derived from work by Traum et al. [24] who use the technique to evaluate the dialogue performance of the virtual humans in search and rescue simulations. They required methods for evaluating not only task success but also the effectiveness of the dialogue given in the situation.

TABLE 2
Appropriateness Annotation Labels and Scores

	Label	Name	Score
User	**RTS**	Response to system	0
	RES	Response received	1
	NRA	No response, appropriate	1
	NRN	No response, NOT appropriate	-2
System	**FP**	Filled pause	0
	RR	Request repair	-0.5
	AP	Appropriate response	2
	AQ	Appropriate question	2
	INI	New initiative	3
	COM	Appropriate continuation	0.5
	NAPE	Inappropriate emotion	-1
	NAPC	Inappropriate content	-1
	NAPF	Inappropriate form, function or other	-1

Webb et al. [25] brought this approach to the evaluation of companions, altering the annotation slightly to suit the domain of companionable dialogues.

The aim is to penalize inappropriate dialogue moves while rewarding those that maintain a natural flow. In order to capture appropriateness of dialogue, each utterance in the dialogue transcript is coded with one of several annotations, shown in Table 2. For user utterances, there are four annotations: user utterances that are a direct response to the system; those that elicit a response from the system; those where no response was received, and this was appropriate behavior; and those where no response was received, and this was deemed inappropriate. For utterances issued by the system, there are nine annotation categories: filled pauses; requests for repair; appropriate responses, questions, new initiatives, and continuations; and finally utterances containing inappropriate uses of emotion or humor, inappropriate *content* of responses (or the content, given the context, of utterances), or inappropriate *form* (or the function of utterances, etc.).

Each of these annotations has a corresponding score, relating the labels to intuitions and understanding of good conversational behavior. For example, appropriate responses and questions from a system are very good (**AP**/**AQ**: $+2$), and extended contributions, where the system expands on an existing topic, are good (**COM**: $+0.5$), but even better are new initiatives and responses pushing an off-topic interaction back on track (**INI**: $+3$). Filled pauses ("umm, err") are generally human-like, and good for virtual agents to perform while filling time, but add nothing to the progression of the conversation (score 0). Repairs and clarifications are bad as such (**RR**: -0.5), and so a system that avoids using them will avoid minor penalties, but their use can be mitigated by allowing subsequent appropriate responses. For example, if it takes two dialogue moves to complete a repair (with a combined score of -1) that then leads to an appropriate response (score $+2$), this subpart of the interaction still gets an overall score of $+1$. This highlights an important aspect of appropriateness annotation, that we can use it to examine subparts of a dialogue, or use it as a reward mechanism in some later dialogue learning strategy. Finally, inappropriate responses of all

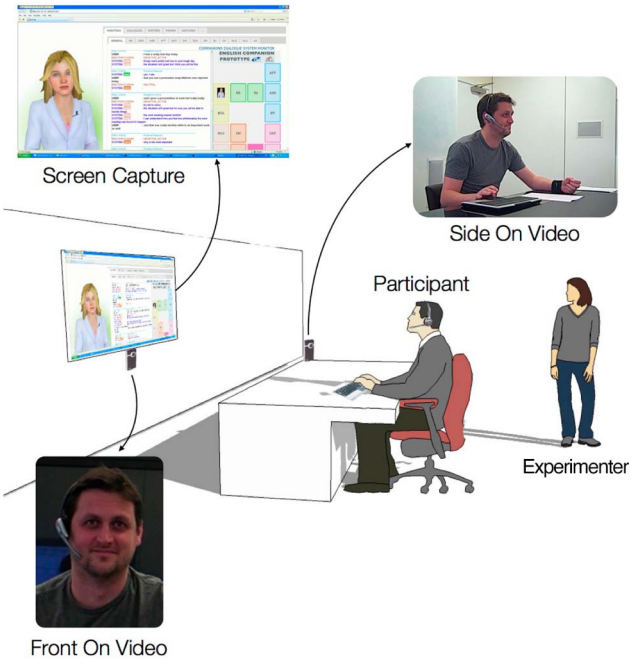


Fig. 2. Evaluation session setup.

kinds (emotion, content or other) are bad (score -1), but no response when one is expected is worse (**NRN**: -2).

Note that these values are presently set by hand, and may not be optimal. At this stage of development, comparative scores and tag distributions across dialogues are more informative, as will be examined further in the evaluation scenarios below.

3 Evaluation Methodology

Using the paradigm outlined in Section 2, the "How Was Your Day" companion was exposed to a number of participants, to test functionality aspects of the complete system. In all, 12 users had a total of 84 separate, fully logged, and recorded formal interactions with the companion in the interactive collaborative environment at Edinburgh Napier University. Participants sat at a desk and faced a 42" LCD screen displaying the prototype interface. Audio-visual recordings were made of each session. Fig. 2 gives a graphical overview of the evaluation layout.

3.1 Participants and Data

All 12 participants were recruited from staff and students at Edinburgh Napier University. Four had some prior familiarity with the COMPANIONS project; the other eight were completely new to it, although some had prior experience with affective or interactive computer systems. Three of the participants were female and nine male; their ages ranged from 22 to 54 with an average of 33. All were native speakers of British English. Users were rewarded for their participation. After the sessions, the participants completed an online user metric questionnaire. For each session, the following data were collected: HD video of each participant (using both front on and side on cameras), video of post session participant interview, prototype screen capture, audio of prototype system, galvanic skin response (GSR; not used in this study since the output was too inconsistent), XML log file detailing all module outputs, and questionnaire response.[2]

3.2 Participant Session Protocol

The following is a description of the session protocol used with each participant of the companion prototype when executing the HWYD dialogue session. Each session took approximately 2.5-3 hours to complete:

1. *Introduction*. The participant was greeted by an evaluator and asked to watch a short video introducing the research, the prototype, the data collection equipment, and the scenario they were to undertake including EmoVoice and ASR training. After the introduction, the participant was asked to sign a video waiver and experiment participant agreement (in line with IRB/ethical treatment of human subjects).
2. *EmoVoice session*. The participant read a short overview of EmoVoice's functionality and was shown a video of someone training on the system to illustrate that the more emotive the users were with their utterances, the more accurate the emotional condition allocation of EmoVoice was. The participant then undertook a training session consisting of reading aloud 42 statements in each of the five emotional conditions (negative active and passive, neutral, positive active and passive). The 210 statements were provided by the EmoVoice developers and are the standard stimulus for EmoVoice training.
3. *ASR training*. The participant went through a Dragon naturally speaking new user training session, in order to provide the ASR model for the prototype.
4. *Prototype session*. The participant was reminded of the scenarios they would be undertaking with the prototype, asked whether they had any questions, and then the session commenced. Participants were instructed not to engage with the experimenter during each scenario stage and to focus attention only on the prototype itself. All recording equipment was activated and the prototype loaded. The experimenter remained in the rear of the room out of the participant's eye line in order to copy the output logs to a server and to restart the system between scenarios.
5. *Postsession questionnaire and interview*. After all scenarios were completed, the participant filled out a Likert scale online questionnaire, and then interviewed for 5-10 minutes on their likes and dislikes of the prototype, the concept, and anything else that came to their mind regarding their experience. Participants were given a reward voucher and thanked.

4 Scenario Design and Scripts

Each participant evaluation session consisted of a set of user scenarios based around templates provided by the system

2. All generated evaluation data (audio, video, affective) is available at the following URL: http://www.napier.ac.uk/companions.

developers, outlining the areas in which the companion was capable of discussing. We designed a set of scenarios to best evaluate the performance of the prototype under certain experimental conditions.

4.1 Pilot Study

In an initial pilot phase members of the evaluation team interacted with the companion, assessing strengths and weaknesses. A total of around 20 scenario combinations were developed to represent the breadth of interaction experience offered by the HWYD scenario. Each scenario session involved a variety of conditions.

A subsequent round of pilot tests of the scenarios led to further refinements: the selection of combinations of utterance types, emotions types, and dialogue initiative that would both evaluate the HWYD system, and show us the value of our evaluation approach.

4.2 Scenarios

With these considerations in mind, six complete scenarios were extracted and the evaluation team refined the scripts to be used for user testing. The scripts were designed to guide the domain of conversation while incorporating enough flexibility for the user to apply their own language choice and to ensure the dialogues were varied. Explicit emotional indicators were provided in each script to ensure the participants were clear on the prescribed emotional state that was intended to guide their language choices and how they would emote, although the choice of, for example, lexical items was left to the user.

In addition to the six scenarios using the prototype UI as provided, it was agreed that an additional interaction session would be undertaken with each participant, only showing the avatar and excluding any other UI elements such as the dialogues in text form. Each scenario contains the following:

1. A feature set:
   a. length of utterance *(short-long-mixed)*,
   b. emotions *(negative-positive-mixed)*,
   c. number of events *(few-many)*,
   d. emotional state *(constant-variety)*.
2. Rationale for using the features (for evaluators).
3. A script guiding the user during conversation.

In most of the scenarios, events were explicitly indicated to users, along with their polarity (how the user should talk about them, in terms of emotional content) and duration (i.e., the scenarios, and by extension the interaction, were considered complete once the script ends). There are two scenarios that are more open-ended, and without the duration constraint. A summary of the scenarios in terms of the feature sets can be seen in Table 3. (In Scenario 5, all the feature settings were allowed to be user defined.) A consequence of this approach, presenting scenarios to users, is the priming effect of language, where we see the language we use to describe the scenarios reflected back in the dialogues. The rest of this section gives a full breakdown of each of the seven scenarios in turn.

Scenario 1a, negative events. This is the baseline condition for the HWYD companion. We found that the system performed best when presented with "negative" events (events of a negative nature as they affect the user). We chose to present only a few events, and to make the overall utterances shorter (in this context, shorter means only one or two events presented to the system at a time). We kept the emotional state of the user constant over the interaction. This structure of scenario consistently gave the best performance in pilot studies. The following script was used:

TABLE 3
Overview of the Scenario Features

Scenario	Utterances	Emotion	Events	Emo. State
1a	Short	Negative	Few	Constant
1b	Short	Positive	Few	Constant
2	Long	Negative	Many	Constant
3	Short	Neg to Pos	Many	Mixed
4	Short	Negative	Few	Constant
5	User def.	User def.	User def.	User def.
6	Short	Negative	Few	Constant

```
NEG Greet Companion
NEG Had a bad day
NEG My promotion was rejected
NEG Gave a bad presentation
NEG Missed an important deadline
NEG Meeting with Nigel & Paul was a disaster
NEG Boss is very unhappy with my performance
```

An example dialogue between the user (U; here named *David*) and the companion system (S; here called *Matilda*) generated from this scenario could start out:

- U: Morning Matilda.
- S: Good morning David, how was your day?
- U: Pretty awful Matilda, I've had a terrible day.
- S: Please tell me.
- U: Well. My promotion was rejected today.
- U: It all happened after I gave a terrible presentation

Scenario 1b, positive events. The pilot studies showed that overall negative events gave the companion greater leverage. However, we wanted a direct contrast. To that end, a minor variant of Scenario 1a was created, with all events positive ("You've had a good day," etc.). This is the only change from Scenario 1a, so presents a clear and direct comparison.

Scenario 2, long utterances. Designed to explore if the system performance changes with long utterances, and whether it is more or less natural to use long or short utterances; also to see the impact on the dialogue of two to three events per utterance versus a single event. The significant change from Scenario 1a is that users are encouraged to offer more information (concepts) to the system in a single turn. We present concepts to the user in the same way, but ask users to combine those events in any way that seems appropriate to them in utterances. Consequently, the overall number of events was increased. We expected this to result in overall longer dialogues, but an interesting contrast in how the system understands the user (e.g., through a concept error rate increase).

TABLE 4
Dialogue Metrics Averages Overall Scenarios

Scenario	Turns		W/utt		C/utt	WER	CER
	User	Sys	User	Sys	User		
1a	13.60	16.60	8.12	6.97	1.31	0.37	0.31
1b	14.67	16.67	8.31	6.51	1.62	0.33	0.31
2	11.00	12.60	10.00	7.63	2.14	0.44	0.34
3	19.67	26.17	10.07	6.58	1.72	0.36	0.34
4	19.17	20.33	9.57	5.90	1.40	0.35	0.39
5	15.50	13.83	10.11	5.41	1.13	0.40	0.26
6	13.40	15.20	6.30	5.55	1.17	0.35	0.33
Average	**15.29**	**17.34**	**8.92**	**6.36**	**1.50**	**0.37**	**0.33**
Range	7–31	3–38	4–23	1–9.21	0.05–4.57	0.15–0.93	0–0.65

NEG Greet *Companion*
NEG Had a bad day
NEG The traffic was really bad this morning
NEG My computer crashed as I was preparing the presentation today
NEG Missed an important deadline
NEG Gave a bad presentation
NEG Meeting with Nigel & Paul was a disaster
NEG Boss is very unhappy with my performance
NEG and so my promotion was rejected
NEG I lost my special parking space
NEG I will miss out on my Christmas holidays
NEG Jane is always harassing me

Scenario 3, mixed emotional states. To this point, the scenarios used fixed emotional states. Scenario 3 was developed with the specific intention of exploring how the system copes with switched emotional state during a conversation, that is, the display of empathy. Negative to positive gave the best performance during pilot sessions, so was the condition chosen for this scenario. The condition is a test of the performance and integration of EmoVoice, in conjunction with the overall dialogue strategy. To produce the clearest results (indicated from pilot studies), this scenario reverted to using short utterances from the user.

Scenario 4, free-form conversation. Scenarios 1a to 3 are extremely controlled. The next two release those controls as an investigation of user behavior when presented with the system. Of course, neither of these scenarios is representative of completely free-form behavior, as each participant will have executed the previous scenarios prior to these, so is intended to have some primed behavior with respect to the companion. In Scenario 4, we explicitly prime the companion with some information, using a correlate of Scenario 1a, before encouraging the user to engage it in free-form conversation for as long as they wished.

Scenario 5, user-defined. To determine how the system copes with entirely user-defined discussion, users were allowed to talk about "their" day in so much as possible, and set no end point in the interaction. Again, as with Scenario 4 we understand the nature of implicit priming, and prior user interactions with the system act as a mechanism for users to understand, at least in part, system functionality.

Scenario 6, avatar only. As seen in Fig. 1, the HWYD system displays a wealth of information, including the avatar, visual feedback of what the speech recognizer had output, and textual output about to be rendered by the TTS. During pilot sessions there were mixed feelings about this interface, specifically that the user spent too much time looking at the textual information, rather than looking at the avatar. On the other hand, textual system feedback can be a vital aid to understand system performance. For effective comparison, a duplicate of Scenario 1a was created, concealing the interface entirely except for the avatar.

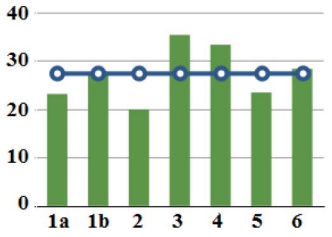


Fig. 3. Average utterance count per scenario (blue line = combined average across all scenarios).

5 Results and Analysis

Twelve participants followed the protocol in Section 3.2, and the setup of Section 3.1 was used to collect three types of data: objective dialogue metrics, emotional speech data from EmoVoice, and appropriateness measurements. These data sets are described in turn below, and the results of the data collection analyzed.

5.1 Objective Dialogue Metrics

Objective dialogue metrics form an important part of any speech system evaluation, and are standardized to some point. We collected a set of metrics (as in Table 1) for each user session: number of turns (user and system), words per utterance (user and system), concepts per utterance (user), word error rate (WER), and concept error rate (CER). Table 4 shows average dialogue metrics scores for all participant sessions and each scenario's average.

5.1.1 Interaction Length

Figs. 3 and 4 demonstrate the relationship between the observable metrics and our a priori beliefs regarding the scenarios. Fig. 3 shows average number of utterances across scenarios, compared to the average across the evaluation. Fig. 4 shows that the average number of user turns was 15.3 and system turns 17.3. Per utterance the average number of words issued by a participant is 8.9, and 6.4 by the companion.

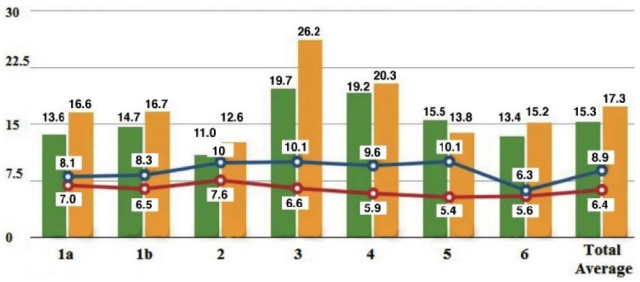


Fig. 4. Average number of dialogue turns per scenario (bars: number of turns; green = user, yellow = system. lines: average words per utterance; blue = user, red = system).

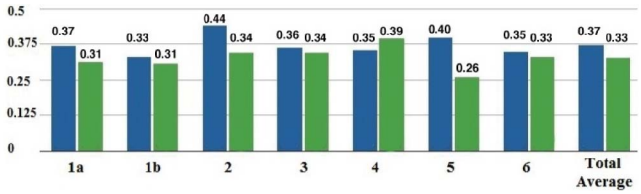


Fig. 5. Average error rates across scenarios (blue = word errors, WER; green = concept errors, CER).

As expected, the shortest interactions are in Scenario 1a using short utterances. Scenario 1b is a very close correlate, and similar in character. Short interactions are also seen in Scenario 2, where longer utterances are used (so taking less interactions to complete the scenario in total), consequently giving less overall utterance count, despite containing more events. Scenario 3 contains mixed emotional content, and prompted longer overall interactions, in part due to the length of the scenario. Scenario 4 is similar initially to Scenario 1a, then allows for a portion of free user input, hence the number of utterances is above average. Interestingly, when users are allowed complete freedom in interaction, as in Scenario 5, the number of utterances drop below average, reflecting the idea that one principal interaction driver was the scenarios. Finally, Scenario 6 is a replica of Scenario 1a, but with reduced visual feedback to the user.

5.1.2 Error Rates

The word error rate was 37 percent on average and the concept error rate 33 percent (Fig. 5). These are very poor scores for speech recognition, and hence present a hard task for any interaction system. The recognizer used was a trainable system, tuned to each participant. However, the system's speech characteristics are tuned to the dictation of prose-type speech, rather than the relatively short utterances seen in dialogues. Furthermore, the requirement on the users to explicitly manipulate their speech to best capture the emotional content of the utterances may have proved a significant downfall in ASR behavior. The worst WER scores were recorded in scenarios where longer utterances were encouraged (e.g., Scenario 2). As expected, CER (although estimated here, as true CER is unknown) is lower than WER. Interestingly, Scenario 5 had the lowest CER at 26 percent, while being the scenario in which the participant was free to discuss any topic they liked. These interactions were shorter in length, and could have been primed by prior scenario performance.

5.1.3 Response Time

In order to establish the average time for the system to respond to a user utterance, the audio waveform from each session was analyzed and the time from the end of user utterance to commencement of audio output from the system was measured. The average response time was 4.18 s (Fig. 6). The evaluators noted whether the audio output came from the short loop or the long loop. When the short loop was activated, the response was at times as low as 1.20 s, with an average of 2.28 s. With long loop responses and more complicated tirades, the average time for response was 6.47 s.

TABLE 5
Results from the EmoVoice Sessions

Actual Emotional Condition	EmoVoice Output: Negative Act.	Negative Pass.	Neutral	Positive Act.	Positive Pass.	Correctly Identified
Negative Active	**251**	22	15	112	62	54.33%
Negative Passive	63	**210**	55	41	93	45.45%
Neutral	41	39	**254**	57	71	54.98%
Positive Active	117	17	42	**197**	89	42.64%
Positive Passive	77	67	51	99	**168**	36.36%
Total	**549**	**355**	**417**	**506**	**483**	**47.67%**

5.2 Emotional Response Analysis

EmoVoice automatically segmented each statement and the next statement was automatically presented to the user. EmoVoice then allocated one of the five emotional conditions to each audio segment.

The scores for 11 participants can be seen in Table 5 (one participant's data were corrupted and lost). As indicated by the last number of the table and the "Total Average" bar in Fig. 7, EmoVoice on average correctly classified 47.67 percent of the statements. It was significantly more successful when identifying negative active (58.92 percent) and neutral (54.98 percent) statements than negative passive (45.45 percent), positive active (42.64 percent), or positive passive (36.36 percent). One possible user influence in this result is that participants typically reported finding it easier to "act" angry or neutral than the other emotional conditions, the passive variants being the hardest. This indicates why we found it expedient to skew evaluation scenarios toward negative events.

Fig. 8 illustrates the emotional condition allocation across all statements by all users (2,310 in total: 42 for each of the 11 users and the five conditions). The EmoVoice results for the participants had a small skew toward negative active, with 23.8 percent of all statements allocated as negative active, and a skew away from negative passive (15.4 percent, versus the actual 20 percent for all).

In order to identify where EmoVoice is allocating incorrect emotional assessments, a similar analysis can be undertaken within a specific emotional condition rather than across all statements. As displayed by the confusion

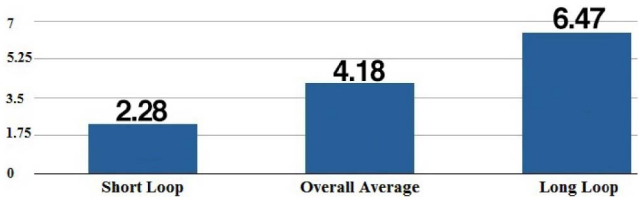


Fig. 6. Average system response time.

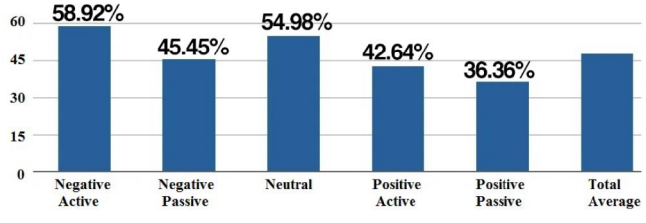


Fig. 7. Average percentage for each emotional condition.

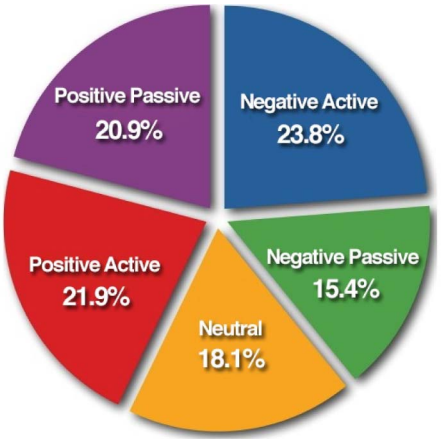


Fig. 8. Emotional condition allocation (in percent).

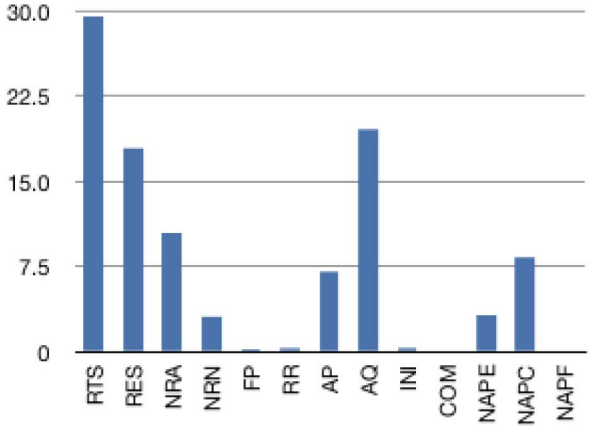


Fig. 9. Annotation distribution (percent) across all dialogues.

matrix in Table 5, for the negative active, negative passive and positive active conditions, the second largest allocation was to the "mirror" emotion: in the negative active condition, it itself had the highest allocation (54 percent, or 251 of 462) and its mirror, positive active, the second highest (24 percent). In the Positive Active condition, 43 percent of the statements were correctly identified, the second highest allocation being the mirror emotion, negative active with 25 percent. In the negative passive condition, 45 percent of the statements were classified correctly, with the mirror emotion, positive passive, being the second most common choice (20 percent).

Interestingly, the one condition in which this did not occur (note, neutral has no mirror emotion) was positive passive, which also had the lowest identification accuracy (36 percent). Here the second highest allocation was to positive active with 21 percent. The mirror emotion, negative passive, was only forth with 15 percent. This may have roots in the "acting" of the participants who found it harder to perform a difference between positive active (e.g., joyful) and positive passive (happy) than negative active (angry) and negative passive (sad). The EmoVoice results reflect that the system had an equally hard time differentiating during the positive passive condition, but more success differentiating during the positive active condition, indicating that EmoVoice is better at detecting more extreme, active emotional states than subtler, passive emotional states.

5.3 Appropriateness Analysis

In conjunction with the objective and subjective analysis performed on most dialogue systems, the component of appropriateness was added. It is a measure of each utterance on a number of dimensions: if it is appropriate given the conversation flow (for example, if a user makes a statement, it may be appropriate to reply, and inappropriate to ignore the speaker); if any use of knowledge in the conversation is handled appropriately (if a user indicates not knowing some persons, it seems inappropriate to ask when they were born); and several other factors, such as the appropriate use of politeness or humor, and error correction strategies that are outside of the present evaluation.

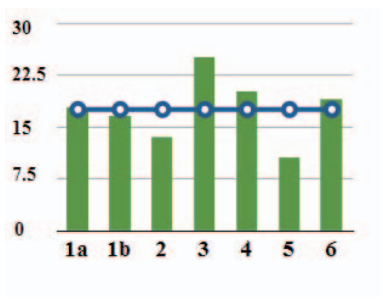
Annotators labeled the appropriateness of every utterance, as described in Section 2.3 (Table 2), given the level of information it contained, and the progression of the dialogue so far. The annotators worked directly with speech recognizer output, and therefore considered the behavior of the system in the context of what the system actually understood.

Three annotators worked on the output of the evaluation sessions. To assess the reliability of transcript labeling, 10 percent of the dialogues were coded for appropriateness by all annotators. Intercoder agreement was calculated using Krippendorff's *alpha* [26] giving an overall agreement of 0.697, sufficient to draw tentative conclusions from the data.

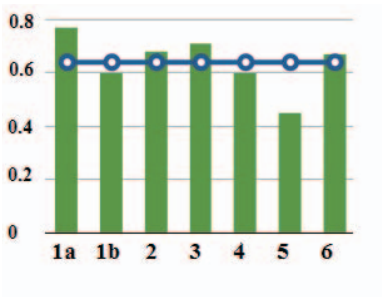
This number seemed low, and we hypothesized that there was significant disagreement between annotators when labeling specific user turns, and visual inspection of the data confirmed this. There was disagreement between annotators when choosing between **RES** (response received) and **NRA** (no response, appropriate) labels. For example, if there was a long series of user utterances followed by a single utterance from the system, annotations were inconsistent as to which of the user utterances resulted in the response (and so should be marked **RES**): the first utterance, the last utterance, or all utterances in the sequence.

Both labels were used when the annotator indicated that the user said something, and that either a response was received, or no response was immediately received to that specific utterance (as opposed to a group of utterances), but that this was acceptable behavior. From an appropriateness annotation standpoint, both contribute the same to the overall score (+1), and so are candidates for conflation into a single category. Recalculating *alpha* with the two categories merged into one results in an increased agreement coefficient of 0.758. A more in-depth exploration of the labels used, in concert with revisions in annotator training, would likely result in further increases in agreement.

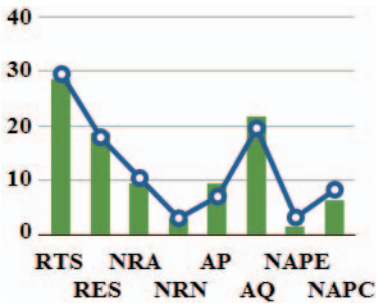
Fig. 9 presents an overview of the distribution of labels across the entire evaluation: the majority of utterances in the evaluation sessions (almost 30 percent overall) are the users' responses to system utterances (**RTS**). The second largest category is appropriate questions asked by the system (**AQ**). Considering the system responses labeled inappropriate, 3.22 percent of the utterances are labeled **NAPE**, i.e., inappropriate as a result of incorrect emotional output (e.g., responding to a negative event with a positive utterance), while 8.31 percent are caused by incorrect semantic content (e.g., a user states that she is working on



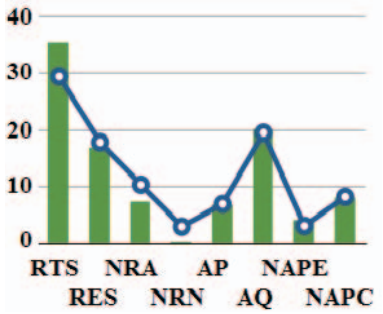
(a) Average score per scenario



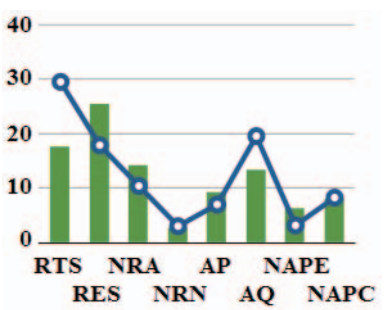
(b) Average score per utterance



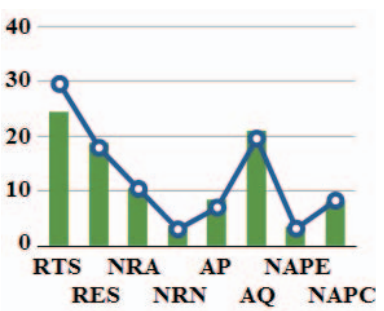
(c) Scenario 1a



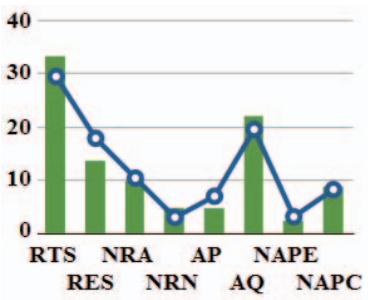
(d) Scenario 1b



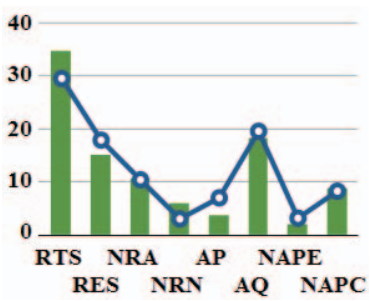
(e) Scenario 2



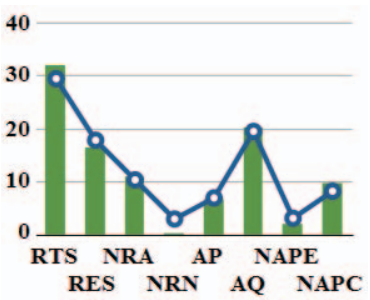
(f) Scenario 3



(g) Scenario 4



(h) Scenario 5



(i) Scenario 6

Fig. 10. Appropriateness scores.

the COMPANIONS project, and the next system question is "What's the name of the project?"). Taking just the inappropriate system responses as a whole, around 30 percent of these errors are caused by inappropriate emotion handling; the remaining 70 percent are from inappropriate content. It appears as though poor performance of EmoVoice has a disproportionately low impact on performance.

The appropriateness annotation can be used to explore each of the scenarios in more detail. First, we compare the performance of the scenarios to the average scores across the evaluation. The average overall appropriateness score for all dialogues is 17.56. Average total score is directly relative to length of dialogue; Fig 10a shows that average score per scenario is also related to dialogue length. The chosen benchmark, Scenario 1a, scores exactly on the overall system average. Scenario 3 is significantly higher (but has significantly higher total utterances), and Scenario 2 is significantly lower (for the inverse reason). What is interesting are the particularly low scores in Scenario 5, the free-form scenario.

Normalizing the appropriateness scores for length of dialogue and showing scores per utterance across scenarios gives the results of Fig. 10b. Here the baseline condition, Scenario 1a outperforms the average, being a very clean and concise interaction. Scenario 1b, by comparison, underperforms the average, despite the only difference being the polarity of events. Most noticeably, scenarios involving any deviation from the script (Scenario 4 with slight deviation, and Scenario 5 with no script) score lower than average.

It is most useful to examine the scenarios in terms of annotation label distributions, and compare to the average scores across the entire evaluation. Figs. 10c, 10d, 10e, 10f, 10g, 10h, and 10i give the distribution of major labels across each scenario, compared to the combined average (the blue lines). By major labels, we mean those showing variance across scenarios, so excluding labels for filled pauses, requests for repair, initiatives, and continuations, as those remain more or less constant.

Fig. 10c shows the baseline condition, Scenario 1a, where the label distribution highly correlates with the average, reinforcing the assumption of this scenario potentially being one of the best performing overall. In Scenario 1b (Fig. 10d), there is larger number of responses to the system, as users give more information in response to systems questions. In addition, while Scenario 1a had few inappropriate emotional responses (**NAPE**), the number in Scenario 1b is above

average: the system struggled significantly more to recognize positive emotional events than negative events.

The Scenario 2 label distribution differs significantly from the previous two (see Fig. 10e). The number of Responses To System (**RTS**) is way below the average, as participants use longer utterances. As a consequence of receiving more information in the utterances, the system ask fewer questions (**AQ** is below average) and the user gives longer, more involved responses to single questions (**RES** is high). A trade-off is that emotional response is harder, resulting in a greater than average number of inappropriate emotional responses: perhaps it is harder to detect the overall emotional value than in shorter, clearer utterances.

Fig. 10f shows the label distribution for Scenario 3, which involved mixed emotional content. The scenario maintained a near average label distribution, where we might have expected a greater number of inappropriate emotional outputs. This could be explained by the overall lack of accuracy of the EmoVoice component across our evaluation.

Scenario 4 represents the first scenario where free-form user input is permissible, following a short script similar to Scenario 1a. Thus, Fig. 10g displays a similar distribution to that in Fig. 10c: the system continues to ask some appropriate questions and the user responds. A slight increase in inappropriate content (**NAPC**, not recognizing the information exchanged from user to system) is also observed.

Scenario 5, where users have complete free access to the system, although guided by prior interactions, gave a change in the relational distribution of three labels. Encouragingly, there is no significant increase in inappropriate responses. However, as Fig. 10h shows, there is an increase in utterances from the user that appear to warrant some response from the system, yet return nothing (**NRN**, where the system is silent in response to some question or emotional comment from the user). We also see a corresponding drop in appropriate responses, and fewer appropriate questions, all of which cause a drop in overall score. As the users deviate from the scripts, the system has less to discuss that is within the topic of the conversation. Consequently, it appears the system chooses to stay silent. Using the simple conversational mechanisms found in chatbots may help to address these issues.

Finally, Scenario 6 with an avatar-only UI (Fig. 10i) shows little deviation from Scenario 1a with avatar plus visual feedback. This scenario was designed to test the UI, and shows that the users and system performed more or less equally, if the user had access to visual feedback from the system or not. In conjunction with the subjective user feedback, this indicates that the visual user feedback should be removed for future trials and use.

5.4 User Experience

The participants completed a Likert scale-based questionnaire following their evaluation sessions with the companion. The questionnaire was composed of 33 statements designed to explore the user experience, and participants answered each by agreeing or disagreeing with each statement. The questions were organized around the six themes described in Section 2.2, which were developed following several empirical investigations of companion technologies. These themes, in conjunction with the objective metrics, allow us to assess the behavior of the companion. Each participant also provided age, gender, and self-assessed rating of their ability to use computer technologies. The scoring for each response was as follows:

Strongly Agree	$+2$
Agree	$+1$
Undecided	0
Disagree	-1
Strongly Disagree	-2

The participants' responses to the questionnaire indicated that they felt the conversation with the companion to be somewhat unnatural. It was not that they thought this to necessarily be an inappropriate thing to do, but rather that it was novel. During the interviews several participants noted that combining the HWYD type discussion with the additional utility of scheduling appointments, say, could prove both cathartic and very useful. Although the participants felt the companion was nothing like themselves (the question "The companion is rather like me," scored -1.6 on average), the participants clearly felt that the companion did have a personality and that it acted independently and demonstrated emotions, specifically that it was polite, friendly, and patient. This personality was felt to be the case despite the occasional lack of coherence in some of the companion's responses and assignment of incorrect emotions to user utterances.

Interestingly, the participants reported feeling that the companion understood them better when there was no textual feedback of either ASR result or emotion detection, as was the case in Scenario 6. They also highlighted that the entire interaction felt far more natural as they focused on the avatar itself rather than the written response as they had intuitively done in the previous sessions. Confusion over turn taking still occurred, but people would spontaneously stop speaking when the ECA started to respond. A linked issue reported by every participant was the lack of communication to the user by the system as to its *internal state*, specifically whether it was or was not *thinking* about what to respond (i.e., was still in a "listening" state) and whether it was going to respond or not. This is a fairly typical usability issue within any computational system where user frustration is increased not by the specifics of UI feedback or the time to receive that feedback, but by the lack of knowing whether any feedback is actually going to come or not.

6 Discussion and Conclusion

We have presented the evaluation of a novel piece of software that was developed to explore the concept of an artificial companion. The companion engaged in interactions with people through an on-screen avatar conducting conversations about how their day at work had been. In 2013, this is still a novel form of interaction. There are only a few ECAs such as GRETA [27] able to engage in these sorts of interactions, and the commercial state-of-the-art Apple Siri, only deals with focused commands with a few "canned" responses to more general questions. We expect companion technologies to develop rapidly over the next few years. In developing the evaluation paradigm for the

HWYD companion, we also make a larger contribution to the evaluation of companions in general.

One limitation of the study was the ability of the companion to recognize speech which, coupled with a relatively slow response time, led to a poorer user experience than could have been the case. Until these issues are fixed, system conversations are obviously always going to be suboptimal. However, we can see from the appropriateness scores that the underlying DM is performing relatively well in selecting responses that people thought were appropriate. Another limitation of the study was that we did not seek to perform a component analysis, although some components require particular attention. In particular, the overall high ASR Word Error Rate hampers many efforts to create companionable dialogues. Given this, the system performed reasonably well, although it has no particular strategies for managing speech error.

Furthermore, the study did not investigate longitudinal interaction with the companion, as the version that we had was not robust enough to be deployed in people's homes or without the support of a research assistant. The sampling of participants for the study thus also had to be restricted to people that would tolerate a lengthy laboratory evaluation. The study hence involved 12 participants, all affiliated to one single university, but with different academic positions. This does not invalidate the results, but it should be clear that the sample might only partially be representative for those intended to use the system.

At the stage of development of the HWYD companion, a scenario-based laboratory evaluation was appropriate. The scenarios themselves were developed to test particular features of the companion's performance on negative and positive dialogues of different types. Thus, the evaluation can be criticized for the lack of ecological validity. The study was performed in a laboratory environment and the next step would obviously be to test the companion with a more longitudinal approach placed in a natural workplace or everyday setting. However, we do not feel that this either compromises the results obtained or the overall evaluation method.

The evaluation scenarios (Section 4.2) were chosen to test specific conditions of the HWYD companion and were able to show some performance issues. For example, there was an implicit belief that the system would perform better with long user utterances, but this was shown not to be the case. As with most spoken language systems, shorter (although significantly longer than most task-based systems) focused utterances proved most successful. The appropriateness annotation provides several interesting features when analyzing dialogues. Specific annotation gives developers key insights into areas of system performance that can be addressed at both micro and macro levels. At micro level, a list of utterances can be output from the system (and surrounding context) and be judged to be inappropriate on some level (providing direction for system improvements). At macro level, the graphs of distribution of labels indicate conversation trajectories that can be useful characterizations of both scenarios and systems.

The traditional objective measures of dialogue quality identified in Section 5.1 obviously remain important for companion interaction. The WER and CER obtained in the evaluation are both high. Clearly, if the speech recognizer fails to correctly register what the person says, subsequent interactions will be less effective. The number of turns and words per utterance are also important metrics, though in the case of companions they should not necessarily be minimized. If the goal is to engage the user in conversations, longer utterances may be appropriate in some circumstances whereas shorter utterances with more turn-taking may be better in others. Work on these measures of appropriateness and their relative weighting needs to continue.

An interesting point to note is that in the participant interviews after all sessions, length of delay in response was considered far less an issue than the timing of the response as well as knowing if and when a response was coming. Participants wanted feedback regarding the state of the companion during the response delay: if the companion was going to deliver a response or not (several utterances per dialogue receive no reply). The largest frustration was when they started talking again but the companion proceeded to talk over them.

In more general terms, the evaluation paradigm can be used to look at a broad range of companions. Speech is central to these technologies and it is important to strive for a natural dialogue between a companion and its owner. This can be measured in terms of response time, appropriateness score, and utterance length. A natural dialogue will be mixed initiative, which can be measured by the length of system and user turns. A good dialogue will also demonstrate stickiness as measured in terms of length of the overall interaction and whether people return to interact with the system. The look and behavior of the avatar is a clear part of the overall user experience as emphasized in the face and body modeling of systems such as GRETA [27]. In line with the idea of relational agents [6] and of moving from interaction to relationships [3], the social, emotional, and psychological aspects of the interaction contribute to the user experience and can be measured through questions such as those included in the user questionnaire. Finally, the utility of a companion is an issue. Companions can span the spectrum of utility from being highly useful (providing information, or caring for someone) to being quite non-utilitarian (e.g., like a pet) but still providing important support to individuals: being able to discuss your working day with your companion may be just what you need.

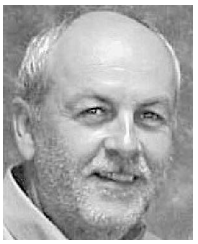
Acknowledgments

This work was partially carried out within the EC/FP6 integrated project companions (IST-34434) while Dr. Webb was at the State University of New York, Albany, and Dr. Hansen and Prof. Gambäck were active at the Swedish Institute of Computer Science AB, Kista. The authors wish to thank Jay Bradley, to the developers of the HWYD companion and of EmoVoice, and to the participants in the user studies at Edinburgh Napier University, Scotland.

References

[1] *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues,* Y. Wilks, ed. John Benjamins, 2010.
[2] Y. Wilks, "Is There Progress on Talking Sensibly to Machines?" *Science,* vol. 318, no. 9, pp. 927-928, 2007.

[3] D. Benyon and O. Mival, "Landscaping Personification Technologies," *Proc. 26th SIGCHI Conf. Human Factors Computing Systems,* pp. 3657-3662, 2008.
[4] R.W. Picard, *Affective Computing.* MIT Press, 1997.
[5] J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation.* Freeman, 1976.
[6] T.W. Bickmore and R.W. Picard, "Establishing and Maintaining Long-Term Human-Computer Relationships," *ACM Trans. Computer-Human Interaction,* vol. 12, no. 2, pp. 293-327, 2005.
[7] A. Nijholt, "Conversational Agents and the Construction of Humorous Acts," *Conversational Informatics: An Eng. Approach,* pp. 21-47, Wiley, 200.
[8] E. André, M. Rehm, W. Minker, and D. Bühler, "Endowing Spoken Language Dialogue Systems with Emotional Intelligence," *Affective Dialogue Systems,* pp. 178-187, Springer, 2004.
[9] A. Paiva, J. Dias, D. Sobral, R. Aylett, S. Woods, L. Hall, and C. Zoll, "Learning by Feeling: Evoking Empathy with Synthetic Characters," *Applied Artificial Intelligence,* vol. 19, nos. 3/4, pp. 235-266, 2005.
[10] C. Smith, N. Crook, S. Dobnik, D. Charlton, J. Boye, S. Pulman, R.S. de la Camara, M. Turunen, D. Benyon, J. Bradley, B. Gambäck, P. Hansen, O. Mival, N. Webb, and M. Cavazza, "Interaction Strategies for an Affective Conversational Agent," *J. Presence: Teleoperators Virtual Environments,* vol. 20, no. 5, pp. 395-411, 2011.
[11] R. Santos de la Camara, M. Turunen, J. Hakulinen, and D. Field, "How Was Your Day? An Architecture for Multimodal ECA Systems," *Proc. 11th Ann. Meeting Special Interest Group Discourse Dialogue,* pp. 47-50, 2010.
[12] T. Vogt, E. André, and N. Bee, "EmoVoice: A Framework for Online Recognition of Emotions from Voice," *Proc. Fourth Workshop Perception Interactive Technologies for Speech-Based Systems,* pp. 88-199, 2008.
[13] K. Moilanen and S. Pulman, "Multi-Entity Sentiment Scoring," *Proc. Seventh Int'l Conf. Recent Advances Natural Language Processing,* pp. 258-263, 2009.
[14] M. Hassenzahl, "User Experience and Experience Design," *Encyclopedia of Human-Computer Interaction,* second ed., ch. 3, M. Søgaard and R. Friis Dam, eds., The Interaction Design Foundation, 2013.
[15] M. Danieli and E. Gerbino, "Metrics for Evaluating Dialogue Strategies in a Spoken Language System," *Proc. AAAI Spring Symp. Empirical Methods in Discourse: Interpretation Generation,* 1995.
[16] W. Minker, "Evaluation Methodologies for Interactive Speech Systems," *Proc. First Int'l Conf. Language Resources Evaluation,* pp. 199-206, 1998.
[17] M.A. Walker, A.I. Rudnicky, J.S. Aberdeen, E.O. Bratt, J.S. Garofolo, H.W. Hastie, A.N. Le, B.L. Pellom, A. Potamianos, R.J. Passonneau, R. Prasad, S. Roukos, G.A. Sanders, S. Seneff, and D. Stallard, "DARPA Communicator Evaluation: Progress from 2000 to 2001," *Proc. Seventh Int'l Conf. Spoken Language Processing,* pp. 273-276, 2002.
[18] M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," *Proc. 35th Ann. Meeting Assoc. Computational Linguistics,* pp. 271-280, 1997.
[19] M. Hajdinjak and F. Mihelič, "The PARADISE Evaluation Framework: Issues and Findings," *Computational Linguistics,* vol. 32, no. 2, pp. 263-272, 2006.
[20] A. Jönsson and N. Dahlbäck, "Distilling Dialogues—A Method Using Natural Dialogue Corpora for Dialogue Systems Development," *Proc. Sixth Conf. Applied Natural Language Processing,* pp. 44-51, 2000.
[21] E. Gerbino and M. Danieli, "Managing Dialogue in a Continuous Speech Understanding System," *Proc. Third European Conf. Speech Comm. Technology,* pp. 1661-1664, 1993.
[22] A. Simpson and N.M. Fraser, "Black Box and Glass Box Evaluation of the SUNDIAL System," *Proc. Third European Conf. Speech Comm. Technology,* pp. 1423-1426, 1993.
[23] L. Hirschman and H.S. Thompson, "Overview of Evaluation in Speech and Natural Language Processing," *Survey of the State of the Art in Human Language Technology,* ch. 13, R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, eds. Nat'l Science Foundation/European Commission, 1995.
[24] D. Traum, S. Robinson, and J. Stephan, "Evaluation of Multi-Party Virtual Reality Dialogue Interaction," *Proc. Fourth Int'l Conf. Language Resources and Evaluation,* pp. 1699-1702, 2004.
[25] N. Webb, D. Benyon, P. Hansen, and O. Mival, "Evaluating Human-Machine Conversation for Appropriateness," *Proc. Seventh Int'l Conf. Language Resources and Evaluation,* 2010.
[26] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology,* third ed. Sage, 2012.
[27] E. Bevacqua, K. Prepin, R. Niewiadomski, E. de Sevin, and C. Pelachaud, "GRETA: Towards an Interactive Conversational Virtual Companion," *Artificial Companions in Society: Perspectives on the Present and Future,* A. Winfield, ed., pp. 143-156, http://www.academia.edu/2669520/Artificial_Companions_in_Society_Perspectives_ on_the_Present_and_Future, 2010.



David Benyon is currently a professor of human-computer systems, the director of the Centre for Interaction Design, and the faculty director for interdisciplinary research at Edinburgh Napier University, Scotland. He has received funding from several European funding programs and developed a number of novel theoretical ideas on human-computer interaction concerning the sense of self and place in mixed reality environments and also published on semiotics and experientialism applied to new media. He is the author of *Designing with Blends* (MIT Press, 2007) and *Designing Interactive Systems* (Pearson, second ed. 2010).

Björn Gambäck is currently a professor of language technology in the Department of Computer and Information Science, the Norwegian University of Science and Technology, and the head of European Collaborations at SICS, Swedish ICT Research AB. He has been at the Royal Institute of Technology and universities in Saarbrücken, Helsinki, and Addis Ababa. He has managed a dozen national and international projects and published more than 100 scientific papers on subjects such as conversational agents, spoken dialogue translation, system and user evaluation, and machine learning applied to language processing.

Preben Hansen received the PhD degree in information studies and interactive media from Tampere University, Finland. He is currently an assistant professor in the Department of Computer and Systems Sciences, Stockholm University. He works with research questions in the areas of information seeking (IS) and information retrieval (IR), including theoretical models of IS and IR processes, empirical studies of users and the use of interactive information access systems, and collaborative environments.

Oli Mival received BA (Hons) degree in psychology from Edinburgh University and the PhD degree in human-computer interaction from Edinburgh Napier University. He is a senior research fellow and the director of the future interaction network at the Centre for Interaction Design, Edinburgh Napier University, Scotland. His research interests include developing, designing and implementing new forms of interface and interaction experience.

Nick Webb is currently a visiting assistant professor in the Department of Computer Science, Union College, New York. His research encompasses a range of language processing applications, including information extraction, question answering and dialogue systems, computer science education, and social robotics. He was a principal investigator of the National Science Foundation-funded Social Robotics Consortium of the Capital Region, and is co-PI of the Social Robotics Workshop, funded by the National Center for Women and Information Technology.