# How Was Your Day? Evaluating a Conversational Companion

David Benyon * Björn Gambäck † ‡ Preben Hansen ‡ Oli Mival * 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 that lack any clearly defined tasks and goals. The development of this type of Companion technology 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 behaviour, with respect to three parameters: functional ability (i.e., does it do the 'right' thing), content (i.e., does it respond appropriately to the semantic context), and emotional behaviour (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, ASR performance and overall system performance (as is expected in systems of this type) but beyond this, we show where individual utterances or responses, which are indicated as positive or negative, show an immediate response from the user, and demonstrate how our combination evaluation approach highlights issues (both positive and negative) in the Companion system's interaction behaviour.

## 1 Introduction

Pervasive, multi-modal conversational systems showing Companionable behaviour present a range of new challenges to dialogue system development and evaluation. In order to be a proper Companion to the user, the system should be able to engage in dialogues lacking both specific tasks and clearly defined goals — except for maintaining the conversation and keeping the user 'satisfied' [1]. Companions differ from traditional dialogue systems in that the conversation is not goal-oriented; however, they are also more than chatbots: a proper Companion must be able to show an appropriate level of understanding of user utterances and respond accordingly. To be truly engaging, such a system should attempt to interpret the emotional state of the user and in turn itself be able to show empathy and possibly even display humour.

Evaluation of such complex, collaborative dialogue systems is a difficult task. Traditionally, developers have relied on subjective user feedback and parameterisation over observable metrics. However, both models place some reliance on the notion of a task; that is, the system is helping the user achieve some clearly defined goal, such as book a flight or complete a banking transaction. It is not clear that such metrics are as useful when dealing with a system that has a more complex task, or no definable task at all.

The paper discusses the use of objective measures, subjective measures and appropriateness annotation for evaluating Companions, and general requirements and features of the approach. We evaluate such a system, the "How Was Your Day" (HWYD) Companion [2], [3], an embodied conversational agent that can discuss work-related issues. In addition to looking at traditional measures such as length of the interaction, we evaluate the HWYD Companion's emotional capabilities, and investigate the use of appropriateness as a measure of conversation quality, the hypothesis being that good Companions need to be good conversational partners.

This introduction 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 discusses the evaluation methodology and 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 analysed 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 [4] is illustrated in Figure 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. The system both

* *Centre for Interaction Design, Napier University, Edinburgh, Scotland.* *E-mail:* {d.benyon,o.mival}@napier.ac.uk
† *Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.* *E-mail:* gamback@idi.ntnu.no
‡ *SICS, Swedish Institute of Computer Science AB, Kista, Sweden.* *E-mail:* {gamback,preben}@sics.se
◇ *Department of Computer Science, Union College, Schenectady, New York, USA. E-mail:* nwebb@union.edu


Fig. 1: The "How Was Your Day" Companion interface

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.

The system exhibits two distinct processing loops 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 which 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. Part of such a conversation between the user and system can be seen in the middle of Figure 1.

Nuance's Dragon Naturally Speaking™ provides the Automatic Speech Recognition (ASR); the recognized words are passed to Dialogue Act Tagging (DAT) which along with information from the acoustic analysis and Acoustic Turn Taking (ATT) allow the system to identify the dialogue acts that are passed to Natural Language Understanding (NLU).

Two modules analyse the emotional content of user utterances: an emotional speech recogniser, EmoVoice [5] returns information indicating the arousal and valence of the acoustic properties of the user's speech as negative-passive, negative-active, neutral, positive-active or positive-passive, while a text-based Sentiment Analyser (SA) [6] operates on the utterance transcript from the ASR, compositionally classifying linguistic units of various syntactic types (noun phrases, clauses, sentences, etc.). It is able to assign 'strength' of the sentiment expressed, but the current implementation simply classifies clauses as negative, neutral or positive. The two emotional inputs are fused together by Emotion Modelling (EM) whose purpose is to provide an aggregate emotional category to be attached to the event description template produced by the NLU and DM. The mechanism for affective fusion overrides the valence category of EmoVoice with the one obtained by SA if EmoVoice's confidence score is below a pre-set threshold value (depending on the competing valence categories).

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) and handed to Natural Language Generator (NLG). The NLG output is passed both 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 module guiding the movements of the avatar, producing gestures and facial expressions conveying the Companion's emotional state.

Two more modules are shown in Figure 1: the Knowledge Base (KB) acts as the central repository of data in the system and is available to all other modules, while the Interruption Manager (IM) [7] handles the system's responses to user barge-ins. When a genuine user interruption (rather than just a backchannel) is detected, the IM instructs the Companion to stop speaking (at next natural stopping point) and the user's interruption utterance is processed by the long loop.

### 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. Some tasks are not defined in such a way that an automatic system can know *a priori* when they are complete. It may be that the task itself is defined as maintaining a relationship, not something that can be measured using traditional metrics such as *task completion*. When devising an evaluation paradigm for such systems, we need to balance the completion of any tasks with some measure of "conversational performance". The assumption in traditional dialogue evaluation is that the quality of the conversation correlates with *user satisfaction*. That is, if the resulting dialogue is annoying or repetitive, we expect a corresponding drop in user satisfaction. However, user satisfaction is in some sense a composite score, covering the entire interaction. Thus can, for example, poor text-to-speech performance have a disproportional effect on user satisfaction.

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. [8], [9], [10]). 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. Note that these are somewhat functional descriptors of quality of interaction.

The DARPA Communicator Program made extensive use of the PARADISE metric [15]. PARADISE (PARAdigm for DIaLogue System Evaluation) was developed to evaluate the performance of spoken dialogue systems, in a way de-coupled 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 *how* it gets accomplished, in terms of the quality measures indicated above. PARADISE aims to maximise task completion, whilst simultaneously minimising dialogue costs, measured as both objective efficiency of the dialogue (length, measured in total turns for example) 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 [16]. For example, using explicit confirmation increases the likelihood of task completion, and so is often chosen, despite being regarded as somewhat unnatural in comparative human-human speech data.

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 develop 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 (cf. [11], [12], [13], [14]).

## 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 behaviour (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:

**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 natural language understanding, in conjunction with readily computable scores such as dialogue duration; number of turns; words per turn, etc.

**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.

| 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 |

TABLE 1: Objective Metrics

Human labelers assign categories to both system and user utterances, with particular focus on system behaviour. 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 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

The 16 objective metrics are outlined in Table 1. Standard timing information needs to be collected from each interaction. Delay times between utterances, both system and user, should be captured, as well as overall dialogue length, in time and in number of utterances. Vocabulary sizes and utterance lengths (in words) are expected to be available both based on ASR results and on transcriptions. Word error rate (WER) is calculated using the standard formula ($WER = \frac{deletions+insertions+substitutions}{number\ of\ words\ uttered\ by\ user}$). Regular dynamic programming string alignment is used to calculate the errors. Concept Error Rate (CER) is calculated by ignoring the order of recognised concepts, where substitution errors are used only for cases where part of the recognised and actual concepts match.

### 2.2 Subjective Measures

Traditional dialogue systems place a high reliance on user feedback. Measures of how people relate to Companions are collected through on-line questionnaires. The questions are organised around six themes that have been developed following several empirical investigations of Companion technologies. The themes all contribute to people developing a sense of social presence of technologies. This encourages people to move from simply interacting with a system to forming a relationship with the technology, which is something that Benyon and Mival [17] have argued is central to the notion of Companions. The themes are:

- **A** Naturalness of the Companion
- **B** Utility of the Companion
- **C** Participant-Companion relationship nature

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

TABLE 2: Measure of appropriateness

**D** Emotion demonstrated by the Companion
**E** Personality of the Companion
**F** Social attitudes of the Companion

These themes, in conjunction with the objective metrics, allow us to assess the behaviour of the Companion as a conversational agent. Some of the themes are geared toward specific behaviours of the Companion system, for example, targeted questions on the use of emotion (both recognizing emotion from the user, and generating appropriate emotion in response to the user) by the Companion. These questionnaires were administered to users following an evaluation session.

### 2.3 Measure of Appropriateness

Appropriateness is a measure of each utterance made by the system, where human annotators score the level of appropriateness given the utterance's level of information and the progression of the dialogue. We principally explore the application of appropriateness as described by Traum *et al.* [14]. The measure of appropriateness penalises mechanisms seen as inappropriate between humans, such as over-verification; strong, one-sided initiative; repetitive behaviour; and the presentation of limited choices, even when these factors contribute to better speech recognition results.

In order to capture appropriateness of dialogue, annotation of the dialogue transcript is required. Annotators used a system splitting the system and user utterances and coded each with one of several annotations, shown in Table 2. For users, 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 behaviour; and those where no response was received, and this was deemed inappropriate. For system utterances, there are nine categories: filled pauses; requests for repair; appropriate responses, questions, new initiatives, and continuations; and finally utterances containing inappropriate uses of emotion or humour, inappropriate *content* of responses (or the content, given the context, of utterances), or inappropriate *form* (or the function of utterances, etc.).

Each of the resulting annotations over the transcript is then scored. First, filled pauses are graded as generally human-like, and good for virtual agents to perform, but do not add a lot (score 0). Appropriate responses and questions are very good (AP/AQ: +2), and extended contributions are good (COM: +0.5), but even better are new initiatives and responses pushing the interaction back on track (INI: +3). Repairs and clarifications are bad as such (RR: -0.5), but their use can still gain points by allowing subsequent appropriate response. 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); thus we still reward this sub-part of the interaction with an overall score of +1. Finally, inappropriate responses of all kinds (emotion, content or other) are bad (score -1), but no response is worse (NRN: -2).

Note that these values are set by hand. When working with such a reward-oriented approach to dialogue modelling in a Companion scenario the measures may be weighted in alternate ways, requiring benchmarking. However, this evaluation methodology can be used to grade complete and part dialogues: the total score (or indeed individual annotation scores) is not necessarily the most useful in all stages of development of a dialogue system. Instead, comparative scores and tag distributions across dialogues can be better measures, as will be examined further 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, twelve 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 and affective data in the form of galvanic skin response was recorded. Figure 2 gives a graphical overview of the evaluation layout.

### 3.1 Participants and Data

The participants were recruited from staff and students at Edinburgh Napier University. Four had some prior familiarity with the Companions project; the remaining 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

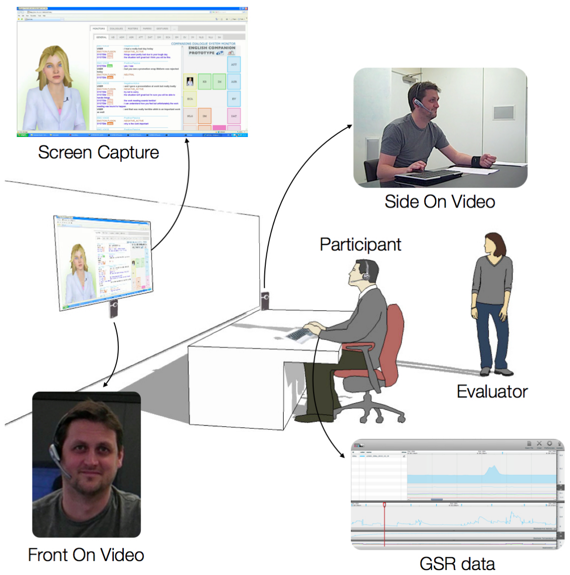



Fig. 2: Overview of the data collection and participant location during each evaluation session

session the participant completed an online user metric questionnaire hosted on surveymonkey.com.

For each session, the following data was collected:

- HD video of each participant (front and side on)
- Video of post session participant interview
- Prototype screen capture
- Audio of prototype system
- Q™ file for Galvanic skin response (GSR) output[1]
- XML log file detailing all module outputs
- Questionnaire response for each participant

All generated evaluation data (audio, video, affective) is available for online access for interested researchers.

### 3.2 Participant Session Protocol

The following is a description of the session protocol used with each participant of the Companions 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 user was, the more accurate the emotional condition allocation of EmoVoice was. The participant then undertook a training session consisting of reading aloud 42 statements for each emotional condition (as detailed in Section 3.3).

**3. ASR Training** Next the participant went through a Dragon Naturally Speaking new user training session, the results of which provided the ASR model for the prototype.

**4. Prototype Session** Once completed the participant was reminded of the scenarios they would be undertaking with the prototype, and to emote as best they could when speaking with the Companion, using the emotional condition as indicated in the scripts for each session. The participants where then asked whether they had any questions, after which the session commenced. All recording equipment was activated and the prototype was loaded. Between each of the scenarios the output logs were copied to an external server and the prototype rebooted.

**5. Post Session 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 then given a reward voucher and thanked. All data was copied to an external drive and collated into a redundant storage array.

### 3.3 EmoVoice Sessions

As was shown in Figure 1, two different modules in the HWYD Companion aim to elicit the emotional content of user utterances: The EmoVoice module [5] analyses the speech input to determine if it is a positive or negative sentiment and an active or passive form, information which the Sentiment Analysis module [6] in parallel tries to elicit from the linguistic data. This information is fused together by Emotion Modelling to a representation of the user's current emotional state in the form of one of five possible values (Negative Active or Passive, Neutral, Positive Active or Passive).

During the evaluation period each participant undertook independent EmoVoice training and testing session in order to examine the accuracy of emotional condition allocation of the EmoVoice system for the users of the prototype system. The participants were given an introduction to the functionality and an overview of how the session would be undertaken.

During each EmoVoice session the participant was asked to read aloud a series of 42 emotionally appropriate statements in each of the five emotional conditions:

- Negative Active: "I really hate how he treats me",
- Negative Passive: "It's got to the stage where I don't care any more",
- Neutral: "Angela Merkel is German Chancellor",
- Positive Active: "I just love to sing and dance",
- Positive Passive: "Today has been a good day".

1. An Affectiva Q Sensor™ from MIT Media Lab measured skin conductance, a form of GSR that grows higher during excitement, attention or anxiety, and lower during boredom or relaxation.

The 210 statements were provided by the EmoVoice developers and are the standard stimulus for EmoVoice training. The participants were asked to "act out" each statement as best they could in the appropriate emotional way, that is, to sound angry if appropriate to the statement; or sad, joyful, neutral, and so on. They were shown a video example of a user undertaking a session to illustrate this. The participants undertook the session in a different room to the Companion evaluation in order to give them some privacy when reading aloud so as best to enable the optimum conditions from emotional allocation by EmoVoice.

| 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 |

TABLE 3: Overview of the scenario features

# 4 Scenario Design and Scripts

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

We conducted an initial pilot phase, where members of the evaluation team exclusively interacted with the Companion, assessing what appeared to be anecdotal strengths and weaknesses. During this initial phase, the evaluation team developed a total of around twenty scenario combinations that best represented the breadth of interaction experience offered by the HWYD scenario. It was decided that this represented too large a set for comprehensive testing, and so these were then scaled down to a design of ten basic scenarios (14 with Positive/Negative variations). Each scenario session involved a variety of conditions.

A subsequent round of pilot tests of the scenarios led to further refinements, including a series of notes that needed to be considered before using the scenarios:

- A user should add information to answer the ECA's questions more appropriately, such as:
  - a project name,
  - a project leader, and
  - people you are working with.
- If and when the ECA takes over the conversation, there is a need to let it lead it.
- Longer user utterances seem more successful.
- Negative events give the ECA more leverage for tirades, whereas overly positive user dialogues offer the Companion little to converse about.

## 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 whilst 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 user interface 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 set of features:
   - length of utterance *(short – long – mixed)*
   - emotions *(negative – positive – mixed)*
   - number of events *(few – many)*
   - emotional state *(constant – variety)*
2) Rationale for using the features (for evaluators).
3) A script guiding the user during the conversation. In most of the scenarios, we were explicit about events, their polarity (how the user should talk about them, in terms of emotional content), and duration (that is, scenarios — and by extension the interaction — was considered complete once the script ends). There are two scenarios which are more open-ended, and do not have this 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.) 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 effect 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:

```
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 be:

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 first thing this morning ...

**Scenario 1b, Positive events:** In pilot studies, we found that overall negative events gave the Companion greater leverage. However, we wanted a direct contrast. To that end, we created a minor variant of Scenario 1a, where all the events were positive. This is the only change from the previous scenario, so would present us with a clear and direct comparison. Script:

```
POS Greet Companion
POS You've had a good day
POS You've been offered a promotion
POS Gave a good presentation
POS Made an important deliverable deadline
POS Had a great meeting with Nigel & Paul
POS Boss is happy with your work
```

**Scenario 2, Long utterances:** This scenario was 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. It was also intended to see the impact on the dialogue of two or three events per utterance versus a single event. In this scenario, the significant change from Scenario 1a is that users are encouraged to offer more information (more concepts) to the system in a single user turn. As a consequence, we had to increase the overall number of events. We expected the outcome from this condition to be overall longer dialogues, but an interesting contrast in how the system understands the user (through a potential concept error rate increase, for example).

```
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 empathy. Negative to positive gave better performance during pilot sessions than positive to negative, so this was the condition we chose to use in this scenario. This condition is a test of the performance and integration of the EmoVoice component, 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.

```
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 Gave a bad presentation
NEG Missed an important deadline
NEG I must work over the Christmas holidays
POS Meeting with Nigel & Paul went very well
POS My promotion was accepted
POS Boss is very happy with my performance
POS I will have extra holidays this year
POS Jane always says how good my work is
POS I was given a special parking space
```

**Scenario 4, Free-form conversation:** Scenarios 1a–3 are extremely controlled. The next two release those controls as an investigation of user behaviour when presented with the system. Of course, neither of these scenarios is representative of completely free-form behaviour, as each participant will have executed the previous scenarios prior to these, so is intended to have some primed behaviour 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.

```
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
BEGIN FREEFORM on any topic the user desires
```

**Scenario 5, User-defined:** In order to determine how the system copes with entirely user-defined discussion, we allowed users 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 Figure 1, the HWYD system displays a wealth of information, including the avatar, visual feedback of what the speech recogniser had output, and textual output about to be rendered by the TTS. During pilot sessions there were mixed feelings about this interface, specifically

| 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 |

TABLE 4: Dialogue metrics averages over all scenarios

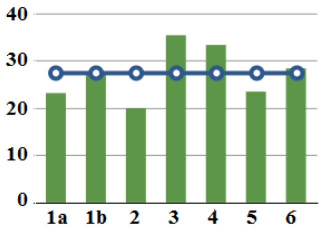



Fig. 3: Average utterance count per scenario
(blue line = combined average across all scenarios)

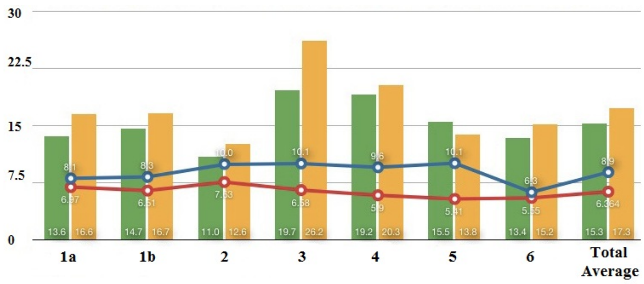



Fig. 4: Average number of dialogue turns per scenario
(bars: number of turns; green=user, yellow=system.
lines: average words per utterance; blue=user, red=system)

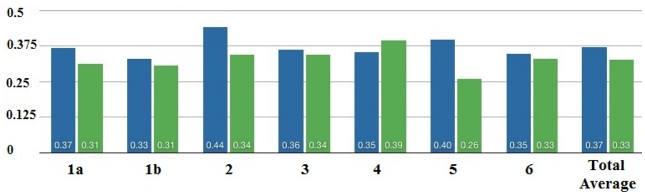



Fig. 5: Average WER and CER across scenarios

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.

## 5 Results and Analysis

Twelve participants followed the Protocol in Section 3.2 and the set-up 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 analysed.

### 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) covering the extent of the scenario dialogues captured during 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*

Figures 3 and 4 demonstrate several of the hypotheses adopted with our evaluation scenarios. Figure 3 shows average number of utterances across scenarios, compared to the average across the evaluation (blue line). The right-most bars of Figure 4 show 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.

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, so is marginally longer than 1a; hence the number of utterances is above average. Interestingly, when users are allowed complete freedom in interaction, as in Scenario 5, the total number of utterances drop below average. Finally, Scenario 6 is a replica of Scenario 1a, but with reduced visual feedback to the user.

#### *5.1.2 Error Rates*

As shown in Figure 5, the word error rate was 37% on average and concept error rate 33%. These represent very poor scores for speech recognition, and hence present a hard task for any interaction voice system. It is difficult to hypothesise why the ASR scores are so low. The recogniser used was a trainable system, tuned to each participant. However, the speech characteristics

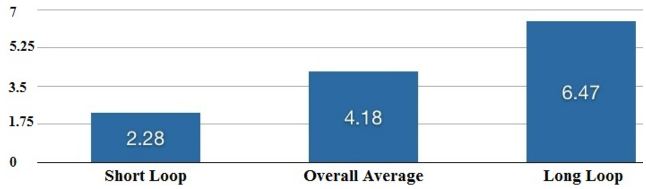



Fig. 6: Average system response time

of this system are tuned to dictation of prose-type speech, rather than the relatively short utterance forms seen in dialogues. In addition, the added overhead of requiring users to explicitly manipulate their speech to best capture the emotional content of the utterances may have proved a significant downfall in the behaviour of the ASR. Thus the worst WER scores were recorded in scenarios where longer utterances were encouraged, as in Scenario 2. As expected, concept error rate (although estimated here, as true CER is unknown) is lower than WER. Interestingly, Scenario 5 had the lowest CER at 26%, whilst being the free-form scenario in which the participant was free to discuss any topic they liked, which in our estimation demonstrates a level of robustness of the system when dealing with concepts outside its core topics.

### 5.1.3 *Response Time*

In order to establish the average time it took for the system to respond to a user utterance, the audio waveform from each session was analysed and the time from the end of user utterance to commencement of the audio output from the system was measured. Typically the user interface would output the text response before the audio output began (to the order of 0.3–1.0 seconds). However, for the purpose of this analysis, response time reflects the audio input-output of the system. The average time from end of user utterance to response was 4.18 s (Figure 6). During the annotation of the waveforms, 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 (ignoring short loop responses), the average time for response was 6.47 s.

## 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 session would take approximately 45 minutes to complete. After each session the evaluators copied the resulting output from EmoVoice into a spreadsheet allowing the assessment of percentage of correct identification in each emotional condition, the breakdown of emotion allocation in each condition, and a total correct identification average.

| Emotion Condition | Negative | | Neutral | Positive | | Correct Identification |
|---|---|---|---|---|---|---|
| | Act | Pass | | Act | Pass | |
| Negative Active | **251** | 22 | 15 | 112 | 62 | 58.92% |
| 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%** |

TABLE 5: Results from EmoVoice session

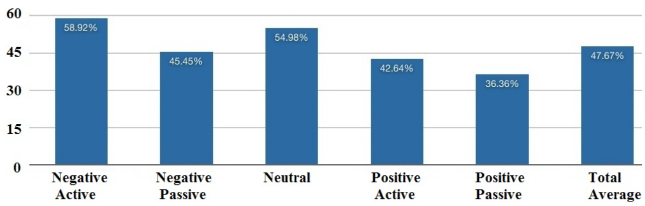



Fig. 7: Average percentage for each emotional condition

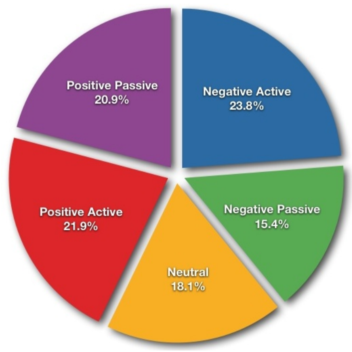



Fig. 8: Emotional condition allocation (in %)

The scores for eleven participants can be seen in Table 5 (one participant's data was corrupted and lost). As indicated by the last number of the table and the 'Total Average' bar in Figure 7, EmoVoice on average correctly classified 47.67% of the statements. It was significantly more successful when identifying Negative Active (58.92%) and Neutral (54.98%) statements than Negative Passive (45.45%), Positive Active (42.64%) or Positive Passive (36.36%). 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 towards negative events.

Figure 8 illustrates the emotional condition allocation across all statements by all users. The EmoVoice results for the participants had a small skew towards Negative Active, with 23.8% of all statements allocated as Negative Active versus the actual 20%, and a skew away from Negative Passive (15.4% versus 20%).

In order to identify where EmoVoice is allocating incorrect emotional assessments, a similar analysis can be undertaken within a specific emotional condition, as in Figure 9, rather than across all statements. For the Negative Active, Negative Passive and Positive Active

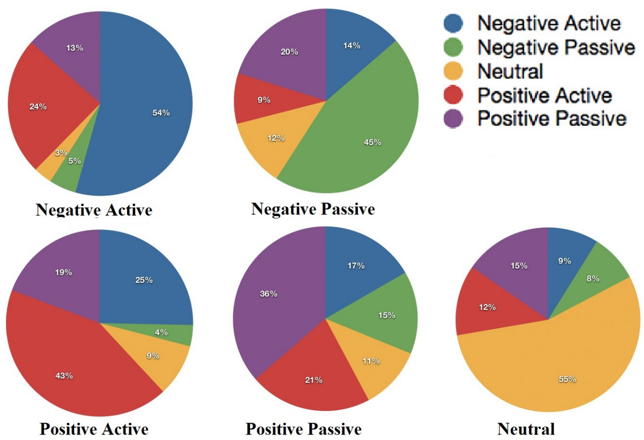



Fig. 9: Emotional allocation division (%)

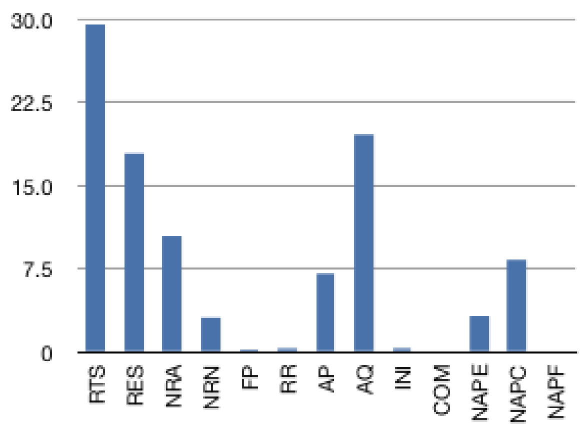



Fig. 10: Annotation distribution (%) across all dialogues

conditions, the second largest percentage allocation was to the "mirror" emotion, i.e., in the Negative Active condition, it itself had the highest percentage allocation (54%) and its mirror, Positive Active, the second highest (24%). In the Positive Active condition, 43% of the statements were correctly identified, the second highest allocation being the mirror emotion, Negative Active with 25%. In the Negative Passive condition, 45% of the statements were classified correctly, with the mirror emotion, Positive Passive, being the second most common choice (20% of the statements).

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%). Here the second highest allocation was to Positive Active with 21%. The mirror emotion, Negative Active, was only forth with 15%. This result may again have roots in the "acting" of the participants who reported that they found it harder to perform a difference between Positive Active (e.g., joyful, ecstatic) and Positive Passive (e.g., happy, content) than Negative Active (e.g., angry) and Negative Passive (e.g., sad). The EmoVoice results seem to reflect that the system had an equally hard time differentiating during the Positive Passive condition, although it had more success with the same differentiation during the Positive Active condition. This indicates 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. Appropriateness is a measure of each utterance on a number of dimensions. Firstly, if it is appropriate given the conversation flow (if a user says hello, it may be appropriate to reply, and inappropriate to ignore the speaker). Second, is any use of knowledge in the conversation handled appropriately (if a user indicates not knowing some persons in a picture, it seems inappropriate to ask when they were born). Third, there may be other factors to consider, such as the appropriate use of politeness, humour or error correction strategies that are outside of the present evaluation.

To conduct the evaluation, annotators scored the level of appropriateness for every utterance, given the level of information it contained, and the progression of the dialogue so far. We want to reward appropriate behaviour (answering questions, using new knowledge correctly) and penalize mechanisms seen as inappropriate between humans: incorrect use of knowledge; asking unrelated or off-topic questions; over-verification; strong, one-sided initiative; and limited choices.

When working with the output of an automatic speech recognizer (ASR), it is necessary to account for that there often is a large discrepancy between what a user actually says and what the system recognizes. The annotations are based on what is recognized **only** — so that if there were recognition errors, the hope would be that either the user spots them in subsequent conversation and can work with the system to correct this, or that the errors are minor in relation to the dialogue flow and hence essentially can be ignored. The system can only function with the content that has been recognised, rather than working on the assumption of completely correct and error-free ASR.

Annotators use a system that splits the *system* and *user* utterances and codes each with one of several annotations, as described in Section 2.3. Three annotators worked on the output of the evaluation sessions. 10% of the dialogues were annotated by all three annotators; pair-wise comparison between annotators on these dialogues shows agreement rates in excess of 90%.

To start the analysis, Figure 10 presents an overview of the distribution of labels across the entire evaluation. A quick breakdown shows that the majority of utterances in the evaluation sessions (almost 30% overall) are responses by the user to system utterances (RTS).

Unsurprisingly, the second largest category is appropriate questions asked by the system (AQ). If we look at the system responses labeled as inappropriate, 3.22% 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), and that 8.31% are caused by incorrect semantic content (e.g., a user states that she is working on 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% of these errors are caused by inappropriate emotion handling; the remaining 70% are from inappropriate content.

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, calculated using the scoring system discussed earlier (see Table 2). Again as noted, average total score is directly relative to length of dialogue; Figure 11a shows that average score per scenario is also related to dialogue length. The chosen benchmark, Scenario 1a scores exactly on the overall system average. Most scenarios are at or above the 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.

Normalising the appropriateness scores for length of dialogue and showing scores per utterance across scenarios, gives the results of Figure 11b. Here the baseline condition, 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 these scenarios in terms of annotation label distributions, and compare them to the average scores across the entire evaluation. Figures 11c through 11i, 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 the scenarios, so excluding the labels for Filled Pauses, Requests for Repair, Initiatives, and Continuations, as these remain more or less constant across all scenarios.

In Figure 11c, we see our baseline condition, Scenario 1a, and observe that the label distribution in this scenario highly correlates with the average. This reinforces our assumption about this scenario potentially being one of the best performing overall.

In Scenario 1b (Figure 11d) there is larger number of responses to the system, as users give more information in response to systems questions. Also, where Scenario 1a had very few inappropriate emotional responses (NAPE), the number in Scenario 1b is above average: the system struggled significantly more to recognize positive emotional events (represented in this scenario) than negative events (Scenario 1a).

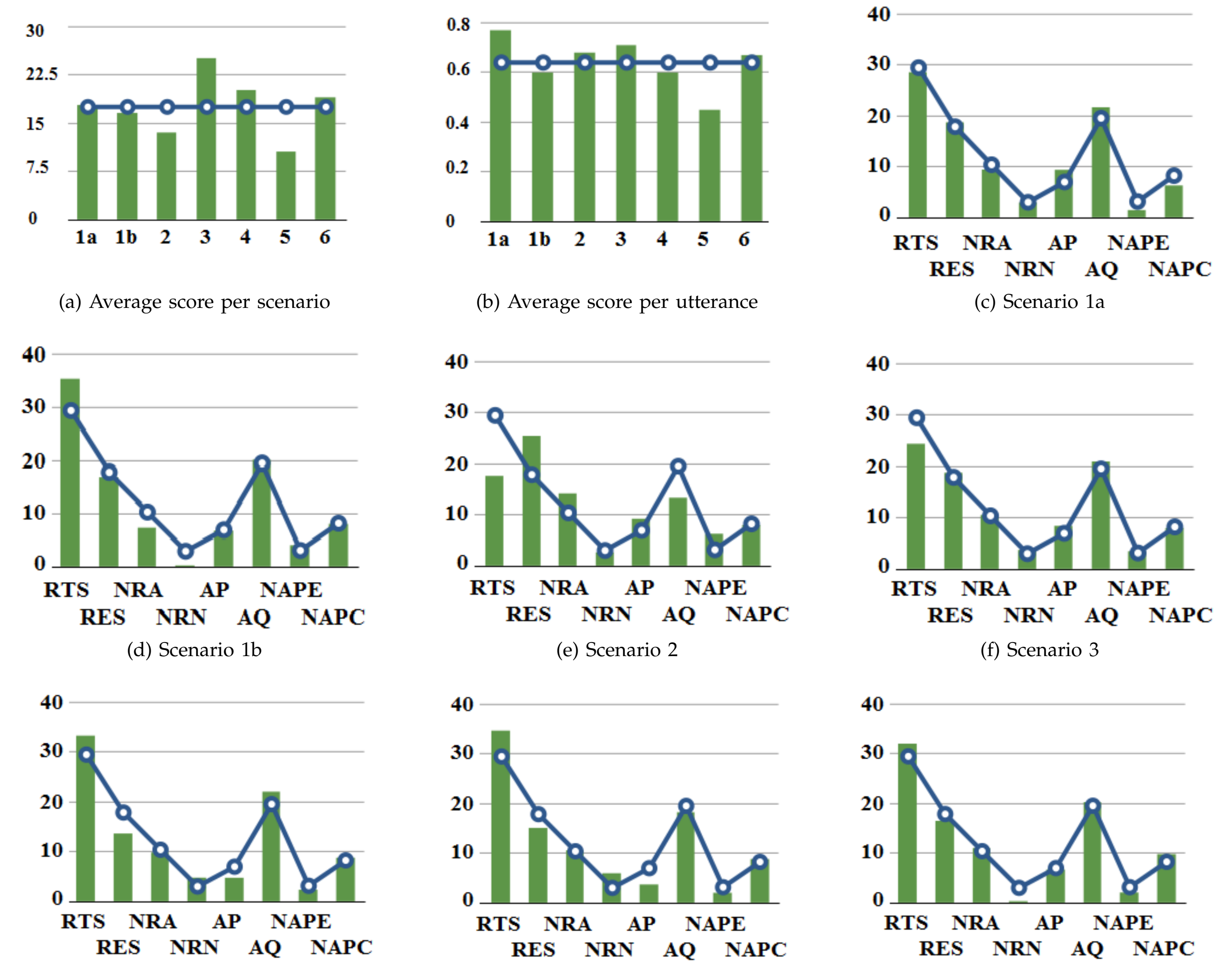
The Scenario 2 (Figure 11e) label distribution differs significantly from the previous two. The number of responses to system (RTS) is way below the average, as participants use longer utterances. As a consequence of receiving more information in the utterances, the system ask fewer questions (AQ is 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.

Figure 11f shows the label distribution for Scenario 3, which involved mixed emotional content. Interestingly, it shows average scores across the scenario for label distribution, where we might have expected a greater number of inappropriate emotional outputs. Given the overall lack of accuracy of the EmoVoice component across our evaluation, we feel that any potential error revealed by this scenario is concealed beneath the general errors of the emotion classification system.

Scenario 4 represents the first scenario where free-form user input is permissible, following a short script similar to Scenario 1a. Thus Figure 11g displays a similar distribution to that in Figure 11c: 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 Figure 11h 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 (and the underlying template structure of the domain) 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 chat-bots may help to address these issues.

Finally, Scenario 6 with an avatar-only user interface (Figure 11i), shows little deviation from Scenario 1a with avatar plus visual feedback (Figure 11c). This scenario was designed to test the user interface, 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 user




(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. 11: Approriateness scores

feedback from subjective surveys, this would indicate that the best course of action is to remove the visual user feedback for future trials and use.

## 6 Discussion and Conclusions

The development of Companion technologies requires new models of evaluation. In this paper, we have concentrated on assessing the HWYD Companion's functionality and overall system behaviour, with respect to three parameters: functional ability (does it do the 'right' thing), content (does it respond appropriately to the semantic context), and emotional behaviour (given the emotional input from the user, does it respond in an emotionally appropriate way).

We have shown how overall system performance, graded on these parameters, is a composite of the lower level system functionality. Equally importantly, we demonstrate the functionality of our evaluation paradigm as a method for both grading current system performance and for targeting areas for particular performance review. We show correlation between, e.g., ASR performance and overall system performance (as is expected in systems of this type) but also where individual utterances or responses, indicated as positive or negative, show an immediate response from the user, and demonstrate how our combination evaluation approach highlights issues (positive and negative) in the HWYD Companion. The evaluation shows that the system performs well, and has an interesting profile when comparing the distribution of appropriateness labels. It is also clear that this represents just a first step towards Companionable dialogue systems. However, the paradigm as deployed gives clear indicators of areas to improve upon.

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 dialogue. Given this, the system performed reasonably well, although it has no particular strategies for managing speech error. Incorporation of these would improve overall scores and feedback. The EmoVoice component may have an effect here. By training for this component, user are effectively shifted from talking in a natural fashion, which directly (and negatively) impacts speech recognition performance. In any case, EmoVoice performance is not ideal, so it is surprising that the system does not output a higher number of inappropriate emotional statements on the basis of this module, possibly since it works in conjunction with a text-based sentiment analysis module, which perhaps mitigates the errors. However, the performance of EmoVoice and the low inappropriate emotion scores correlate with circumstance of WER and CER, that is, one has impact, although not linear, on the other.

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. Participants wanted feedback regarding the *state* of the Companion during the response delay, specifically if the Companion was indeed going to deliver a response or not (there are several utterances per dialogue that receive no reply). They reported that the length of the delay was less impactful than not knowing if and when a response was coming, and the largest frustration was when they started talking again but the Companion then proceeded to talk over them.

The scenarios 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. First, 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. For example, if we want the users to talk more, we need data corresponding to Figure 11e (Scenario 2), where users emit longer utterances. Conversely, if our profile looks more like Figure 11c, we have a more traditional short utterance, interactive dialogue system. Different dialogue strategies may be planned around different dialogue trajectories as indicated by these graphs. Used at the data collection stage, such graphs might present interesting ways to determine optimal system performance, based on user expectation.

If we take the goals of the evaluation paradigm, to develop metrics that can score conversational dialogue systems, the HWYD Companion is successful at achieving some of these 'goals':

**Natural Dialogue:** the user interacts with the artificial agent in a natural way. That is, there are no significant delays in the interaction, the agent uses knowledge in an appropriate way, asks appropriate questions, does not rely on overly strong confirmation strategies, etc. The interactions with the HWYD Companion within domain are mostly appropriate. Out of domain presents a more significant problem, as for most dialogue systems. There are no significant interaction delays, although users indicate that delays are not as important as clarity of signaling turn taking, and the paradigm may be modified on this basis.

**Initiative:** there is a balance between the initiative of the system and the initiative of the user. Either can ask questions, change the topic of conversation, hold the floor if required. Further analysis indicates that the use of appropriateness labels can shed more light on initiative, e.g., at which points in the dialogue is initiative largely given to the user? By plotting initiative over time, an even exchange of initiative as the dialogue progresses should be seen. Again, this may lead to refinements of the evaluation paradigm.

**Confusion:** that the system runs dialogues in a way that does not increase th user's cognitive load. This is the hardest to measure in systems with limited error correction routines incorporated into the dialogue scenario: simple measures of requests for repair can not be used to give some indication of cognitive load.

**Stickiness:** the Companion is desirable to talk to, both within an individual interaction and over a significant period of time (weeks or months). It would be very interesting to evaluate user interaction with the HWYD Companion over a longer period of time.

**User Satisfaction:** the measure of how happy a user is with the interaction, both in the immediacy (at the time of an interaction) and in the long term. The user satisfaction survey results are mixed, and clearly there are component level issues (e.g., speech recognition) which are significant contributors to performance, but it is clear that the sheer novelty of the scenario has a significant impact on user evaluation; users are not yet prepared to hold conversations with computer systems in this way, although it would be interesting to see how users adapt to this scenario over time.

## Acknowledgments

This work was partially carried out within the EC/FP6 integrated project Companions (IST–34434), and while Dr. Webb was at State University of New York; Albany, New York, USA.

Thanks to the developers of the HWYD Companion and the developers of EmoVoice, as well as to Jay Bradley and the participants in the user studies at Napier University, Edinburgh.

## REFERENCES


[1] Y. Wilks, "Is there progress on talking sensibly to machines?" *Science*, vol. 318, no. 9, pp. 927–928, Nov. 2007.

[2] C. Smith, N. Crook, S. Dobnik, D. Charlton, J. Boye, S. Pulman, R. Santos 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," *Presence: Teleoperators and Virtual Environments*, vol. 20, no. 5, pp. 395–411, Oct. 2011.

[3] M. Cavazza, R. Santos de la Camara, M. Turunen, J. Relaño-Gil, J. Hakulinen, N. Crook, and D. Field, "How was your day? an affective companion ECA prototype," in *11th Annual Meeting of SIGDial*. Tokyo, Japan: ACL, Sep. 2010, pp. 277–280.

[4] R. Santos de la Camara, M. Turunen, J. Hakulinen, and D. Field, "How was your day? an architecture for multimodal ECA systems," in *11th Annual Meeting of SIGDial*. Tokyo, Japan: ACL, Sep. 2010, pp. 47–50.

[5] T. Vogt, E. André, and N. Bee, "EmoVoice: A framework for online recognition of emotions from voice," in *4th Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Kloster Irsee, Germany: IEEE, Jun. 2008, pp. 188–199.

[6] K. Moilanen and S. Pulman, "Multi-entity sentiment scoring," in *7th Int. Conf. on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, Sep. 2009, pp. 258–263.

[7] N. Crook, C. Smith, M. Cavazza, S. Pulman, R. Moore, and J. Boye, "Handling user interruptions in an embodied conversational agent," in *9th Int. Conf. on Autonomous Agents and Multiagent Systems*. Toronto, Canada, May 2010, pp. 27–33.

[8] M. Danieli and E. Gerbino, "Metrics for evaluating dialogue strategies in a spoken language system," in *Spring Symposium on Empirical Methods in Discourse: Interpretation & Generation*. Stanford University, California: AAAI, Mar. 1995.

[9] W. Minker, "Evaluation methodologies for interactive speech systems," in *1st Int. Conf. on Language Resources and Evaluation*. Granada, Spain: ELRA, May 1998, pp. 199–206.

[10] 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," in *7th Int. Conf. on Spoken Language Processing*, Denver, Colorado, Sep. 2002, pp. 273–276.

[11] E. Gerbino and M. Danieli, "Managing dialogue in a continuous speech understanding system," in *3rd Eur. Conf. on Speech Communication and Technology*. Berlin, Germany: ESCA, Sep. 1993, pp. 1661–1664.

[12] A. Simpson and N. M. Fraser, "Black box and glass box evaluation of the SUNDIAL system," in *3rd Eur. Conf. on Speech Communication and Technology*. Berlin, Germany: ESCA, Sep. 1993, pp. 1423–1426.

[13] L. Hirschman and H. S. Thompson, "Overview of evaluation in speech and natural language processing," in *Survey of the State of the Art in Human Language Technology*, R. A. Cole, *et al.* Eds. National Science Foundation, European Commission, Nov. 1995, ch. 13.

[14] D. Traum, S. Robinson, and J. Stephan, "Evaluation of multi-party virtual reality dialogue interaction," in *4th Int. Conf. on Language Resources and Evaluation*. Lisbon, Portugal: ELRA, May 2004, pp. 1699–1702.

[15] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain: ACL, Jul. 1997, pp. 271–280.

[16] M. Hajdinjak and F. Mihelič, "The PARADISE evaluation framework: Issues and findings," *Computational Linguistics*, vol. 32, no. 2, pp. 263–272, Jun. 2006.

[17] D. Benyon and O. Mival, "Landscaping personification technologies," in *26th Annual SIGCHI Conf. on Human Factors in Computing Systems*. Florence, Italy: ACM, Apr. 2008, pp. 3657–3662.

[18] H. P. Grice, "Logic and conversation," in *Syntax and Semantics*, P. Cole and J. L. Morgan, Eds. New York, New York: Academic Press, Jun. 1975, vol. 3: Speech Acts, pp. 41–58.


**David Benyon** is Professor of Human-Computer Systems and faculty director for interdisciplinary research at Edinburgh Napier University, Scotland, as well as director of the Centre for Interaction Design at Napier. Prof. Benyon has developed a number of novel theoretical ideas on human-computer interaction concerning the sense of self and sense of place in mixed reality environments. He has also published on semiotics and new media and on applying experientialism to new media. He is the author of *Designing with Blends* (MIT Press, 2007) and *Designing Interactive Systems* (Pearson, 2$^{\mathrm{nd}}$ ed. 2010).

**Björn Gambäck** is Professor of Language Technology at Norwegian University of Science and Technology and Head of European Collaborations at SICS, Swedish Institute of Computer Science AB. He has also worked at the University of the Saarland; Helsinki University; the Royal Institute of Technology, Sweden; and Addis Ababa University. Prof. Gambäck has been Coordinator or Principal Investigator of a dozen national and international projects and has published over 100 scientific papers on subjects such as conversational agents, system evaluation, spoken dialogue translation, and machine learning.

**Preben Hansen** is a senior researcher at SICS, Swedish Institute of Computer Science AB and holds a Ph.D. from Department of Information Studies and Interactive Media, Tampere University, Finland. Dr. Hansen 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 use of interactive information access systems, and collaborative environments.

**Oli Mival** is Senior Research Fellow and Director of the Future Interaction Network at the Centre for Interaction Design, Edinburgh Napier University, Scotland. Dr. Mival holds a degree in Psychology from Edinburgh University and a PhD in Human Computer Interaction from Napier. His research is focused on developing, designing and implementing new forms of interface and interaction experience.

**Nick Webb** is Visiting Assistant Professor in the Department of Computer Science, Union College. Dr. Webb's research encompasses a range of Language Processing applications, including Information Extraction, Question Answering and Dialogue Systems, Social Robotics, and Computer Science Education. He was the Principal Investigator of the NSF-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 (NCWIT).