SEMAINE

THE SENSITIVE AGENT PROJECT

D6d

Evaluation of the Automatic SAL system

Date: 17 December 2010

Dissemination level: Public

ICT project contract no.	211486
Project title	**SEMAINE** **Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression**
Contractual date of delivery	**17 December 20**
Actual date of delivery	*17 December 20*
Deliverable number	D6d
Deliverable title	**Evaluation of the Automatic SAL system**
Type	Report
Number of pages	28
WP contributing to the deliverable	WP 6
Responsible for task	Roddy Cowie (r.cowie@qub.ac.uk), Gary McKeown (g.mckeown@qub.ac.uk) Jennifer Hanratty
Author(s)	Roddy Cowie, Gary McKeown & Jennifer Hanratty
EC Project Officer	Philippe Gelin

Table of contents

1 Executive Summary

Previous deliverables have described the SAL scenario and the system. Evaluation of the system involved the development of suitable measures, a set of five experiments, and assessment of promising lines of development in the light of the data, formal and informal.

Seven measures were selected. Three were assessed by questions asked by a distinct avatar after each interaction with an individual SAL character: *system competence* (how often did you feel the avatar said things completely out of place); *sense of flow* (how naturally do you feel the conversation flowed?); and *engagement* (how much did you feel that you were involved in the conversation?). The engagement measure was supplemented by having a third party use a trace measure to rate the user's *apparent engagement*. A *nonverbal* concurrent measure was provided by a 'yuk button' which users pressed when they felt that the interaction was going badly. The last measure, *simple affect,* was obtained at the end sessions where users experienced two versions of the system (which system did you prefer?). System competence is generally distinct from the others. The two engagement measures, self- and observer-ratings, correlate highly, but the latter seems to be less noisy, and give additional information about change over time. Engagement, flow and nonverbal measures coincide for some characters, but not all. The results summarise below indicate that this battery of measures seems to provide useful information without being unduly intrusive.

The first two experiments used a Wizard of Oz system, 'Semiautomatic SAL', where a human operator chose the system's utterances. There were two types of comparator, one where the human operators had full audiovisual information about the user, and one where they had limited feedback. Results provided information about the measures, and a comparator for ratings of the automatic system.

The third experiment studied the contribution that emotion detection makes to human-avatar interactions. It compared a full version of automatic SAL with one where estimates of the user's emotional state were replaced by random values. No consistent overall effect was found. Taking that in conjunction with findings from the earlier experiments, the likeliest conclusion is that the SAL scripts are unexpectedly robust: they allow operators to function with very limited knowledge about users' emotional states.

The fourth and fifth experiments studied the contribution that emotional expressiveness makes to human-avatar interactions. Each compared a full version of automatic SAL with one where expressive functions (vocal, facial, head nods, etc) were curtailed. In both cases, the full version received significantly better ratings. In the fourth experiment, the effect was due to the negative characters. Recordings were used to modify the positive characters for experiment five. The result was a strong overall advantage for the full system, extending across all of the available measures. Overall average rating of the full system was also higher than either human-operated system in experiment 2 on all the measures that they shared. These findings make it clear that the system does achieve the fundamental SEMAINE goal of building a system that can conduct emotionally coloured interactions with human users.

These global findings are supplemented by more specific findings in several areas. Different SAL characters have different optimum configurations, vindicating SEMAINE's decision to work with multiple characters. User personality affects ratings: it is important either for evaluation or for application that some personalities effectively do not engage with this kind of system at all. There are context effects: response generally improves as people become familiar with the system, but the wrong initial experience reduces engagement. The characteristic of poor experience seems not to be that engagement is not achieved, but that it falls off.

While the studies are partly evaluation of a final system, they are also partly pointers to natural ways of improving the system. Studying the system highlights areas where it seems possible and worthwhile to improve its abilities. Five in particular stand out: knowing when to come in; taking the other party's state as a topic of conversation; 'drawing out' the other party to express views and feelings about a topic; understanding what makes an acceptable sequence of utterances (particularly, but not exclusively, in the course of exchanges that are drawing the other party out or talking about his/her state); and showing emotional flexibility. Working on these areas is not simply about improving SAL: it is about understanding skills that would be potentially useful in a range of applications. What SAL provides is a context where that kind of challenge can be recognised and addressed.

To make the most of that kind of strategy, the system needs to be set up in a way that allows relatively self-contained adjustments to be made, and their effects evaluated. The current version of the system does not allow that. However, because it is fundamentally modular, it would make sense to develop it in that direction.

2 General background and aims

Systems like SAL call for evaluation at several different levels. As a first approximation, lower level issues can be separated out and addressed in comparatively straightforward ways — for instance, by measuring how often emotion is identified correctly from voice alone. This report does not comment on those issues. Its focus is on the less concrete, but at least equally important high-level problem, which is to evaluate the system as a whole.

It is recognised that high-level evaluation of emotion-oriented systems presents particular challenges (Westerman et al 2006). SAL is one of the most purely affective systems in existence (if not the most). Hence evaluating the system is not simply about applying standard methods – it is about developing ways of dealing with various problems that have no agreed solution. Part of the challenge is that there was no way to know in advance what it would be like to hold conversations with an autonomous affective system. As a result, a key part of the evaluation process was identifying relevant variables – dependent variables that expressed the character of the experience, and independent variables that affected it.

One point at least was clear subjectively. At every stage of development, the research team was taken aback by the quality of interaction that people achieved with SAL. In one sense, that is a subjective judgment. However, the grounds for it are publicly available in the form of recordings that make up D6c. Hence the question addressed in the studies reported here was not so much whether the system achieved something interesting and potentially useful, but rather how to pinpoint what was interesting and potentially useful, and capture it 'in black and white'.

This report summarises the main points of the process. In order to do that, it covers early experiments which uncovered important problems as well as the final studies which confirm in a broad sense that SAL's emotional competence enhances users' experience.

3 Bases for selecting measures

Although the literature does not provide ready-made evaluation methods, it offers a variety of pointers which helped to define a reasonable agenda. This section gives a brief overview of them.

The obvious starting point is the substantial literature on usability, which offers well-defined resources. However, it is underpinned by the conception of usability stated explicitly in ISO 9241, which defines it as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." (ISO 1998, p. 2]. Effectiveness and efficiency are not the issue in a system like SAL. Satisfaction has an affective component, but even it is defined in functional terms, as lack of discomfort, and a positive attitude towards the system, while performing the goals. Clearly that does not cover everything that people might look for in an affective system. As Edwardson put it, "We don't ski to be satisfied, we want exhilaration" (1998, p. 2).

More richly affective measurement systems have been developed gradually, but they too often have specific goals in mind. Westerman et al. identified four main areas where measurement has developed: computer anxiety; trust and loyalty; frustration; and 'flow, fun, and playfulness'. The first two appear not to be relevant for SAL. Frustration clearly is, and one might assume it was simply undesirable. However, that is not necessarily so. There is a kind of frustration that is a mark of human engagement. If we treat them as people, then it is right and proper that we should be frustrated by Obadiah's relentless pessimism or Poppy's relentless brightness. Engagement is also a key issue in the last area. If Spike is convincing, an encounter with him is neither fun nor playful. However, it does create the characteristic 'flow' feeling of being engrossed in a task, to the exclusion of distractions (see Csikszentmihalyi & Csikszentmihalyi, 1975).

Highlighting engagement points to another cognate area. One of the issues is clearly whether users feel intuitively that they are engaged in a real conversation with a real personality, and respond realistically, despite knowing intellectually that the other party is not real. These are closely related to the issues that have been highlighted in research on presence (e.g. Sanchez-Vives and Slater 2005); and, of course, they are also related to the Turing Test (Schroeder & McKeown 2010).

A final point that recurs in the literature is scepticism about verbal measures (Ibister et al 2006). There are two main problems associated with them. One is that they disrupt the interaction that they are supposed to be describing. The other is that verbalising memories of emotion-related experiences involves quite radical transformations (Pennebaker & Chung, 2007), so that the reports systematically misrepresent the experiences.

On that basis, we identified six broad types of measure that it seemed appropriate to consider.

Competence
Traditional evaluations focus on system competence. In the case of SAL, the scope for that kind of measure is limited. However, there is one obvious competence to consider, that is, conversational competence – whether the avatar appears to say appropriate things at the right time.

Sense of flow
From the literature, probably the most promising measure is the sense of 'flow', which reflects a judgment that people make easily and reliably, even though its meaning may be hard to articulate.

Engagement
This is related to flow, but there is a difference which is conceptually important. Engagement is specifically a state of the user, whereas flow is a characteristic of the activity – which is this case is jointly constructed by the user and the system.

Simple affect
There are ways of measuring like and dislike which have the potential to be very misleading in this context – asking 'do you like Spike?' misses the point. However, measures at a more general level may well make sense. It seems plausible that people may be able to judge whether they like interacting with a system irrespective of whether they like the characters.

Non-verbal measures
The literature does not suggest any obvious non-verbal measures, but logically it seems natural to ask for a response when the interaction is going badly. That turns an apparent problem to advantage. Giving the signal is a secondary task. Secondary tasks are routinely used as measures of engagement, because people who are immersed in the primary task tend to forget them. Hence it would be inappropriate to ask for a response when the interaction was going well, because the user would probably not be attending to the task when the response was appropriate. However, if the task is to make a secondary response when the interaction is going badly, then the times when attention is likely to be available are precisely those when it is needed.

'Emotional Turing'
In principle, it is natural to consider Turing-like tests, where the question is whether users can tell whether the system is autonomous or at least partly controlled by a human operator. One can imagine arranging situations where there was genuine uncertainty about the question. That avenue was not pursued at this stage, for two main reasons. One is that it would become very elaborate, involving arrangements that would allow a human operator to control various aspects of the system's behaviour. The other, deeper reason is that the question is fundamentally a distraction. It is not the aim of the project to create avatars that will deceive people into thinking that they are human. Among other things, that would be ethically unacceptable. To put resources into passing Turing-like tests would be to let the undoubted fascination of the test take priority over the project's natural (and perfectly ethical) goals.

On the basis of these considerations, we developed and explored a collection of techniques that seem to be suited to this particular evaluation task. They were developed through a series of pilot experiments.

4 Evaluations using Semi-automatic SAL

Semi-automatic SAL is a system where the operator is a human, who chooses from a predefined script what the system's next utterance should be, and when it should be spoken. The actual sound, though, is the pre-recorded voice of an actor who had been chosen to suit the SAL character who was in play at the time. The system's output is purely audio. Visually, all that it offers the user is a schematic face, attached to a display that varies with the sound: its function is simply to provide a natural focus of attention.

The main motivation for the Semi-automatic SAL studies was to provide benchmarks against which the automatic system could be compared. The intention was to obtain different levels of performance by manipulating feedback to the human operator, so that ratings of the automatic SAL system could be located relative to ratings of various semi-automatic systems: at one extreme, a semi-automatic system controlled by a human operator with full feedback; and at the other, a system controlled by a human operator trying to conduct a conversation with virtually no information about what the operator was saying.

Various secondary motives also existed. One was to check that SAL-type interactions could be conducted with the kind of limited linguistic information that an automatic SAL system has. Another was to test the evaluation system.

4.1 Experiment 1

Experiment 1 was a relatively small study designed mainly to estimate the scale of differences produced by varying feedback to the operator, and to confirm that basic measures are sensitive enough to capture them. It is reported because the results were the first indication of an unexpected pattern, and as such, it is a significant part of the evidence relative to that pattern.

Method
The evaluation used here consisted of verbal reports on three key issues – perceived competence of the system; the user's sense of engagement; and the perceived 'flow' of the conversation. It is a known problem that asking for verbal reports during an interaction is likely to disrupt it; and on the other hand, reports given afterwards are likely to rationalise it. The solution developed for SAL was, in effect, a spoken questionnaire designed to minimise disruption by letting users respond from within the scenario. Immediately following each interaction, a different character stepped in and asked (orally) three questions about the interaction that has just finished. The questions target linked, but potentially separable aspects of the interaction:
a) How naturally do you feel the conversation flowed? (targeting user/avatar interaction).
b) How often did you feel the avatar said things completely out of place? (targeting system competence).
c) How much did you feel that you were involved in the conversation? (targeting user state).
Answers to a) and c) were on a scale from 0 (worst) to 10 (best). Answers to b) were in terms of five categories, never (scored as 5); once or twice; a few times; quite often; most of the time (scored as 1). Question b) was placed between the others because since it is in a different format, it reduces the tendency to answer the others in the same way.

Feedback to the operator was varied, on the rationale explained above. In the full feedback condition, the operator could both see the user (via a teleprompter screen) and hear him/her (via loudspeakers). In the mid feedback condition, the visual channel remained, but the auditory channel was filtered by cutting out frequencies between 350Hz and 4000Hz: this gives a reasonable impression

of prosody, but very few words can be made out. In the low feedback condition, the video channel was removed, leaving only the filtered audio. Interactions with Spike always used full feedback. Each participant also interacted with the other three characters, each with a different level of feedback. The different levels of feedback were balanced across the three characters.

Results

Table 1 summarises the results

	Average of Q1	Average of Q2	Average of Q3
Low			
Ob	6.25	3.50	7.25
Poppy	6.25	3.75	7.50
Prudence	8.00	4.75	9.00
Mid			
Ob	7.00	4.00	8.50
Poppy	6.00	4.25	7.00
Prudence	8.25	4.00	8.50
Full			
Ob	6.00	3.75	6.50
Poppy	7.25	3.75	9.00
Prudence	6.25	4.25	8.25
Spike	6.25	3.92	7.67

Table 1

The key result is that level of feedback had no robust effect on ratings. Ratings on all the scales remained relatively close, and well within the upper half of the range, even when feedback to the operator was only filtered speech. This was thoroughly unexpected. It is also notable, though less surprising, that responses to the different characters were practically indistinguishable.

The unexpected outcome has one highly positive implication, which is that a SAL conversation can in principle be carried out without understanding the linguistic content of the user's utterances. Hence trying to design a system that can carry out a SAL-type dialogue without understanding the users' words is not asking the impossible.

On the other hand, the outcome signals that the situation has some features that were not foreseen. It is known that people can recover information about emotion from the kind of filtered speech that was used here. Nevertheless, their performance is substantially different from ratings of full audio-visual speech (Cowie, 2009). The expectation had been that operators' ability to detect users' emotion would deteriorate in the low feedback condition; hence, that they would be less able to make appropriate choices of utterance; and hence, that users would react by rating the interaction less favourably. The results indicate that there is at least one flaw in that chain of reasoning, but it is not at all clear where it is.

A final issue is that the results may have reflected a global positive orientation to the characters. There are two reasons why that might be the case. The first is that half of the participants were students doing the study as part of a project, and the other half were their friends. The students regarded the project as an enjoyable one, and could have approached it with a positive bias. The second is that the questions were asked by a synthesised voice which the participants found very disturbing, and that could have made the characters seem very appealing by contrast.

4.2 Experiment 2

This experiment addressed essentially the same issues as the previous one, with modifications to

take account of the unexpected findings. Given the possibility that even filtered speech is unexpectedly informative, feedback was reduced by eliminating sound altogether, but leaving vision. Given the possibility that the evaluation questions are ineffective, additional measures were introduced.

The feedback manipulation was straightforward. For two out of the four characters that each participant spoke to, the sound was switched off so that the operator could not hear what they were saying, and thus rely on this to select an appropriate response. Each conversation with the four characters lasted 20-30 minutes.

The standard questions were presented in the same way as in Experiment 1, but two additional types of measure were introduced.

The first new measure was a non-verbal concurrent task. Users were given a button to hold during the conversation, and were asked to press it whenever they felt that the simulation was not working well. It was christened a 'yuk button', and the name has stuck. This provides a measure of engagement during the interaction. There is a degree of subtlety in it: the more engaged users are, the less likely they are to think of the button-press task, even if they do feel that the interaction is anomalous in some way.

The second modification was that recordings of the interactions were subjected to impressionistic analysis, aimed at identifying behaviours associated with good and poor interactions. Two separate analyses were carried out, one concerned with auditory indicators, the other with visual indicators.

Results

Table 2 shows the broad pattern of the results. The main outcome is to reinforce the findings of experiment 1. Overall scores on all four measures – flow, correctness, engagement, and concurrent task (the 'yuk button') were extremely close for full and degraded systems.

Measure	speaker	**Mean/character**	
		Degraded	Full
flow	Poppy	5.33	5.83
	Spike	3.6	6.25
	Obadiah	5	4
	Prudence	5.6	3.5
	All	4.885	4.89
correctness	Poppy	3.33	4.33
	Spike	2.40	3.75
	Obadiah	3.60	4.00
	Prudence	4.00	3.00
	All	2.67	2.23
engagement	Poppy	4.67	6.17
	Spike	3.8	6.25
	Obadiah	5.6	4.75
	Prudence	6.4	3.5
	All	5.12	5.17

Yuk button	Poppy	0	1
	Spike	1.6	0.5
	Obadiah	1.6	1
	Prudence	0.8	1.75
	All	1	1.06

Table 2

Correlations were used to explore the relationships among the different measures. Results are summarised in table 3. Significant relationships are shown shaded. Two points stand out. First, as might be expected, questions 1 and 3 are strongly correlated. Hence there is a case for dropping one or other. Second, relationships between indicators seem to be character-dependent. For two characters, Spike and Prudence, concurrent indicators and all three verbal measures intercorrelate in a single global evaluation. For the others, verbal probes a) and c) hang together, but b) and the concurrent task seem to be unrelated. The point is a simple one: impression that covary with one character need not covary with another. The implication is that measures need to be wary of assuming that 'one size fits all'. That links to a general point which is clearly important, and not raised in the literature: different emotional styles pose different challenges for affective computing.

	q1vsq2	q2vsq3	q1vsq3
obadiah	-0.02	-0.08	0.85
poppy	0.00	0.19	0.87
prudence	-0.95	0.75	0.87
spive	-0.77	0.68	0.88

	q1vsyuk	q2vsyuk	q3vsyuk
obadiah	0.02	0.14	0.29
poppy	0.31	0.29	0.22
prudence	-0.48	-0.55	-0.59
spike	-0.38	-0.36	-0.16

Table 3

Impressionistic analysis adopted a simple strategy. For each participant, flow and engagement scores (which were highly correlated) were used to identify the two best interactions and the two worst. The recordings were then studied to identify features of the user's behaviour that appeared to distinguish the two. Results are not conclusive, but promising. Interactions that were poorly rated on the verbal probes tended to show reduction in a range of behaviours, both visible (looking sideways or down, head movements, and hand gestures) and vocal (long utterances, amused laughs, exclamations); and increases in some vocal ones (nervous laughs, unfilled pauses, short utterances, sighing and audible breathing).

The main point of the impressionistic analysis was to identify signs that might allow automatic analysis to recognise when an interaction was in difficulty. However, the findings add to the credibility of the verbal measures. The point is not simply that there are behavioural differences corresponding to the verbal measures (chance almost guarantees that some such differences will be found if one looks hard enough), but that the behavioural differences make a degree of sense.

4.3 Conclusion

1. The measures appear to be reasonably satisfactory. To some extent flow and engagement appear to be redundant, but there is reason to be wary of trimming too far: there may be significant cases where they diverge.
2. One of the functions of these studies was to provide a benchmark against which the automatic system could be measures. The measures in experiment 1 are probably not suitable because of the particular personal involvement that the participants had. However, the measures from experiment 2 seem to be a reasonable comparison
3. It seems that there are observable correlates of engagement and disengagement. That has two kinds of implication. In the long term, it suggests that automatic detection of user engagement is a real possibility. In the short term, it suggests that observation by a third party could be a useful addition to the repertoire of evaluations.
4. One of the key questions behind these experiments was whether a human operator can do what the SAL system is expected to, and carry out an acceptable interaction with minimal information about the words that the users speaks. It is clear that the answer if affirmative. What is striking, and unexpected, is that performance can be so good with so little information. The implication seems to be that conducting an acceptable SAL interaction may not depend nearly as much as we thought on information about the user's emotion. That casts an interesting light on the SAL scripts. They appear to have found a style of interaction which is quite remarkably robust. The basic reason appears to be that they give users phrases that they can 'play off' if they choose to, so that a creative user can turn even an inappropriate SAL phrase into a springboard for a new turn in the dialogue.

   That robustness has both advantages and disadvantages for SEMAINE. The advantage is that the scripts alone make it reasonably certain that an automatic system will be able to conduct a conversation. The disadvantage is that the effect of the nonverbal skills that SEMAINE has developed are not likely to be as clear cut as was initially expected, because interaction will probably not collapse without them. However, that leaves scope for less dramatic effects. The measures that have been developed should be able to capture those.

5 The core experiments

Three experiments were run. Each had individual aims, and they are described relative to those specific aims. Taking them together produces an overall pattern, and it is described after the individual parts have been presented.

The presentation is selective. In the nature of the work, each experiment generated a large amount of data, and a great many findings. The aim here is not to describe them exhaustively, but to convey the core points that each study adds to our understanding of the interactions with SAL-type agents and the task of evaluating them.

5.1 Procedure in the core experiments

The basic experimental design was that each user interacted with two versions of the system – the most competent version available, and a version with minimal nonverbal skills.

As in experiments 1 and 2, the core evaluation questions were kept within the scenario: they were asked by an avatar between sessions with individual SAL characters. In these experiments, the avatar was the Greta agent in its pre-SEMAINE embodiment. The disturbing voice used previously was replaced by a standard unemotional synthesised voice.

Because of concerns about experimenter effects in experiment 1, a preparatory briefing was scripted and delivered at the start of each session by the same avatar.

In experiments 4 and 5, two additional measures were introduced.

Because users experienced two versions of the system, it was possible to ask them at the end of the session which versions they had preferred.

The evidence from impressionistic analysis suggested that it could be useful to include ratings made by a third party. That was done using a trace technique of the type used elsewhere in SEMAINE. An experienced tracer observed interactions as they took place (audiovisually, but in a separate room), and used a specially designed interface to record the user's apparent level of engagement from moment to moment.

Presentation order was randomised within blocks; which system came first was counterbalanced.

In all cases, the SEMAINE system operated in a distributed configuration on two Windows 7 Dell machines using Intel Core i7-860 2.80Ghz CPUs with 4GB RAM. The system was configured with ActiveMQ and the java components and openSMILE operating on one machine and the Greta and visual components operating on a second machine.

In total 90 participants were tested. 14 of these were in piloting and preparation stages, 9 before experiment 1 and 5 between experiments 2 and 3. 16 participants took part in Experiment 1, 11 female and 5 male. 30 participants took part in Experiment 2, 19 female and 11 male. 31 participants took part in Experiment 3, 25 female and 6 males, one female withdrew due to ill health.

5.2 Experiment 3

The specific aim of experiment 3 follows on from the unexpected pattern in the previous experiments. There is an obvious explanation for the limited impact of reducing feedback to the operator, which is that in the SAL scenario, knowing the user's emotions does not always matter a great deal. That is thoroughly unexpected, but it does make sense. Given the game-like quality of the scenario, interacting with an avatar that completely misread the user's emotions could be as much fun as interacting with one that read them correctly.

The first experimental objective with Automatic SAL was to test the relevance of emotion sensitivity directly. This was done by comparing two versions of the system. One (the 'full' system') was the best available at the time, SEMAINE version 3.0.1 (svn revision 734). The other (the 'degraded') system, was the same except that the output of the emotion detection components was replaced by a random function.

Results

Figure 1 shows the basic results for the competence measure. Analysis of variance showed that the full version was rated significantly higher for competence, with $t(57) = 2.46$, $p = 0.02$. That is to be expected given that the degraded version was likely to choose utterances like 'you shouldn't be so cheerful' when it would have been obvious to a human that the user was far from cheerful.

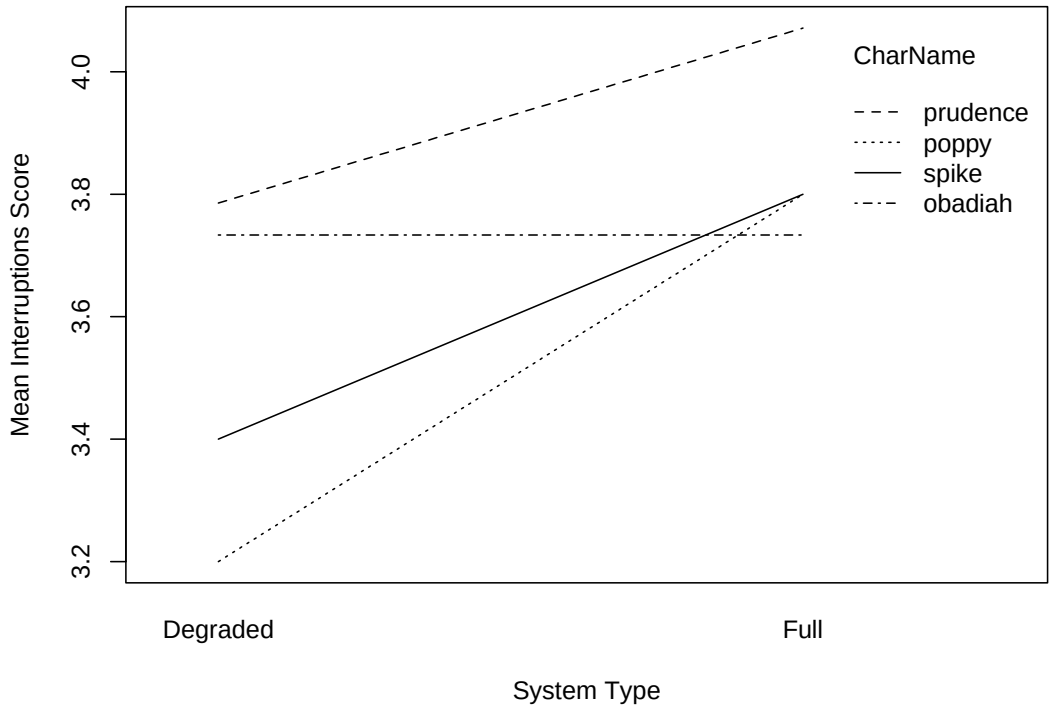


Figure 1

Figure 2, however, tells a different story. The flow measure showed a significant difference between the systems, with $t(59) = -2.46$, $p = 0.017$, but it was the degraded system that users rated higher. That strongly supports the provisional reading of experiments 1 and 2, which is that that in the SAL scenario, knowing the user's emotions matters much less than we had assumed. That conclusion is unexpected and important. There are good grounds to qualify it, and it will be revisited, but the

combined evidence of three studies makes it difficult to doubt that something of that general kind is true.

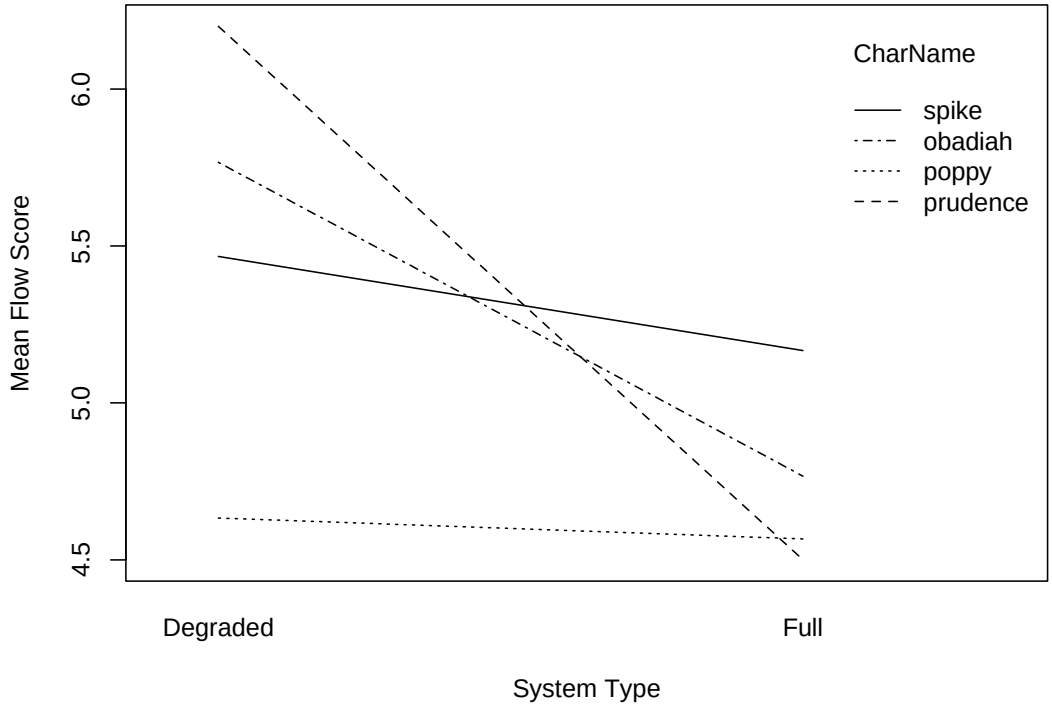


Figure 2

For completeness, the systems were not significantly different in terms of engagement.

Comparisons with experiment 2, where the operator was a human, add interesting points. For correctness, ratings of the full system were very similar to the ratings of the semi-automatic system (3.53 and 3.85 for degraded and full systems respectively, as against 3.33 and 3.77 in experiment 2). For flow, rating of the full system (4.75) was similar to ratings of both variants in experiment 2 (4.88 and 4.90 for degraded and full systems respectively), but rating of the degraded system was substantially higher (at 5.52). For engagement, both versions of the system were rated well above those that were given in experiment 2 (6.26 and 6.30 for degraded and full systems respectively, as against 5.12 and 5.17 in experiment 2). Overall, the purely automatic system seems to match the human-operated one in perceived competence, and provide a more interesting subjective experience.

The relatively high ratings of the automatic system are not mysterious. The obvious hypothesis is that they reflect the richer output of the automatic system. In particular, the semiautomatic system has minimal visual output, whereas the automatic one includes multiple visual strands – appropriate faces, lip movements matched to the speech, appropriate emotional expressions, and backchannelling movements. Hence the next experiment moved to test that the various output elements did indeed have a major effect on quality of interaction.

5.3 Experiment 4

The main focus of this experiment was on the effect that automatic SAL's rich output systems have. That led in two directions.

First, observations from the previous experiment were used to identify problems with the 'full system' used there. The most striking was over-eager intervention. Adjustments were made to reduce the likelihood of the system intervening when speakers were pausing briefly for thought or breath, or even simply lowering their voices. A number of other technical problems were also noticed and addressed.

Second, a degraded system was constructed which differed substantially from the new full version in terms of output mechanisms. The voices were manipulated so that the system spoke with flat affect synthesised voices. The Greta components were largely turned off, so that there were no mimicking or backchanneling behaviours: only speech-related lip movements and blinking remained. The dialogue management components were adjusted so that utterances were selected randomly, and that the decisions when to interject with an utterance would be based on a temporal random distribution.

At the same time, we introduced a number of experimental refinements.

- the personality of the user was measured using standard psychometric instruments
- the trace measure of engagement (described above) was introduced
- The 'yuk button' was incorporated into the experimental setup.

Results

As would be expected, the systems were rated similarly on competence (the directly relevant aspects of the system were the same). Ratings of engagement also showed no difference, as in the previous study. However, the full system was rated more positively for flow, with $t(119) = 2.36$, $p = 0.02$. As figure 3 shows, the effect derived from two characters, Spike and Obadiah.

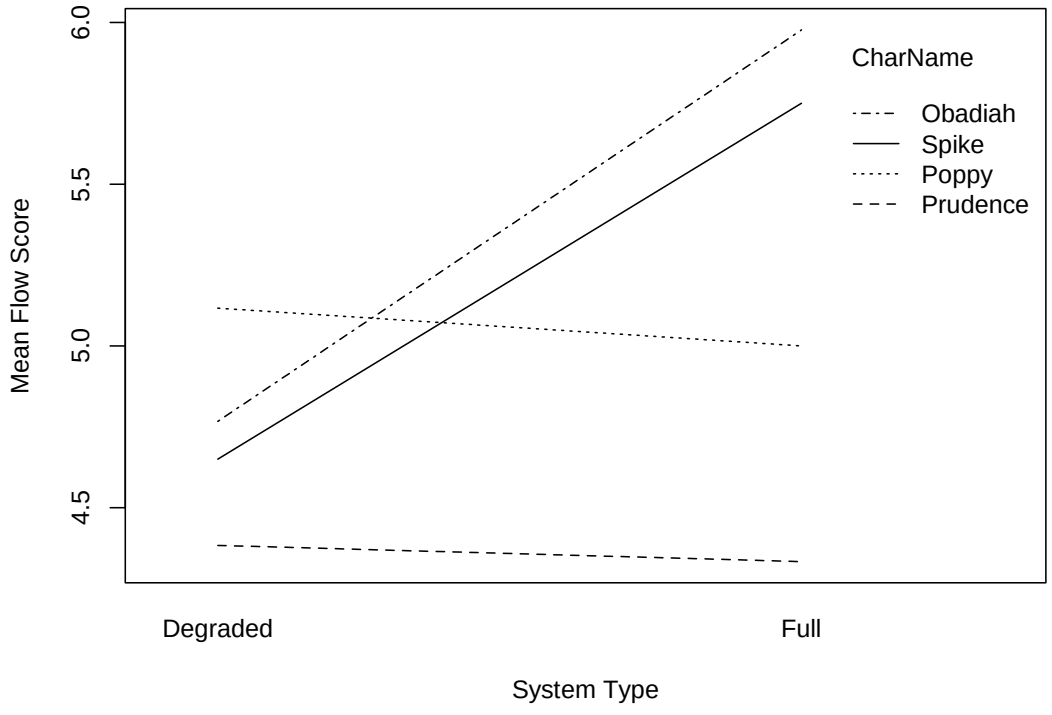


Figure 3

The trace of engagement was introduced partly because we suspected that self-ratings of engagement were subject to individual and situational variables, and felt that an outsider's judgment might be stabler. The trace measures proved interesting on a number of levels.

	Degraded				Full			
	Obad	Poppy	Prud	Spike	Obad	Poppy	Prud	Spike
self-rating	5.95	5.8	5.23	5.77	6.35	5.82	5.52	6.4
trace	0.56	0.53	0.52	0.55	0.63	0.57	0.56	0.64

Table 4

Table 4 makes an important initial point. It shows the mean self-rating of engagement for each character*system combination (on a scale of 0-10), compared to the mean of the engagement trace (on a scale of 0-1). The correspondence is striking. It gives a strong indication that there is a quality that is understood in very similar ways by the person experiencing it and by an outside observer. But because different extraneous variables influence the two kinds of estimate, they may behave very differently as regards statistical analysis.

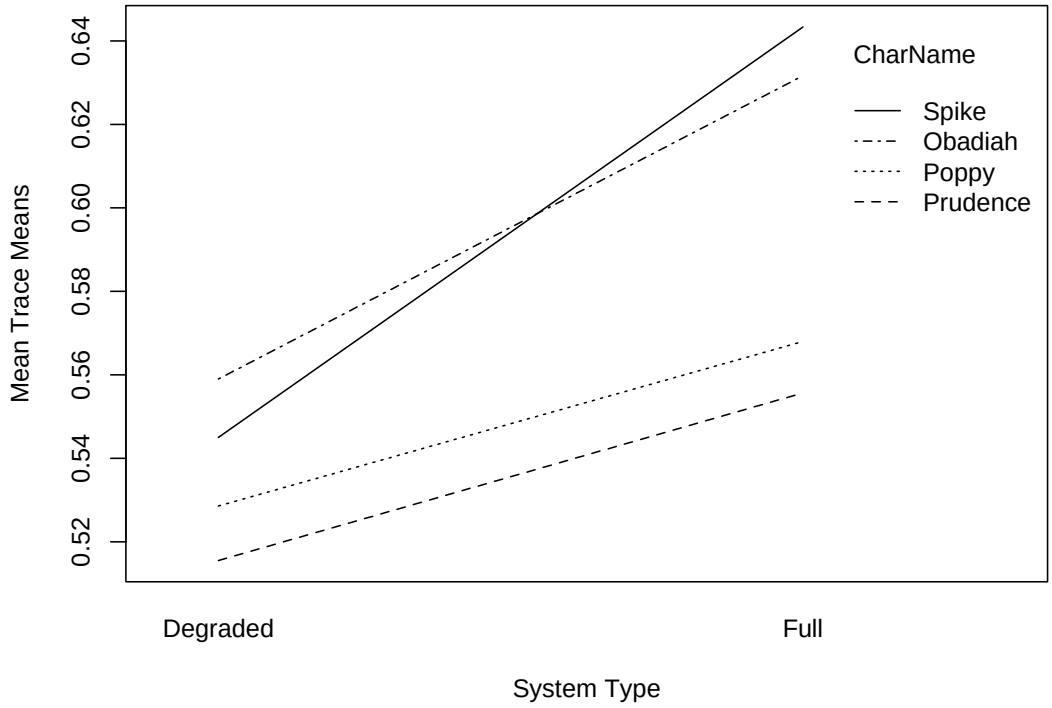


Figure 4

Figure 4 shows the difference in mean engage traces. The difference between systems is highly significant, with $t(116) = 6.80$, $p < 0.001$. Combining Table 4 and Figure 4, it seems quite possible that traces can provide what is essentially a less noisy measure of the same underlying variable as ratings.

Ratings on the trace also allow another level of exploration. We applied the stylisation techniques described in SEMAINE D6b, and found credible patterns relating to the way engagement changed with time. Two key indices of change were much less with the more competent system – the

maximum magnitude of rises (F(1,228)= 9.78, p=0.002) and the minimum magnitude of falls (F(1,229)= 9.79, p=0.002). The median duration of level stretches also showed significant effects, in this case with with character (F(3,229)= 5.54, p=0.001): it was higher with the better received characters, Spike and Obadiah. Correspondingly, the frequency of falls varied with character F(3,229)= 4.52, p=0.004: it was lower with the better received chatacters. In broad terms, what the pattern suggests is that quite varied systems and characters can reach high levels of engagement with: but with less satisfactory systems and characters, it falls away. It is important for the evaluation of practical systems to understand that the challenge is not to reach peaks of engagement, but to keep it there.

Fixed Effects			
	Estimate	Std. Error	t
(Intercept)	5.25	0.68	7.71
Order	0.23	0.05	5.15
Psychoticism	-0.72	0.24	-3.03
System Type	1.15	0.29	3.9
Character Valence (Positive /negative)	0.04	0.29	0.14
Interaction of Character Valence& system	-1.22	0.42	-2.94
Random Effects	Std.Dev		
Participant (Intercept)	1.74		
Residual	1.61		

Table 5 Explanatory Multilevel Model for Flow

The data also allow another level of detail to be clarified. In a complex, subjective task, one would expect experience to be affected by various factors. It is possible to explore these because the number of participants was large, and personality measures were taken for each. Table 5 shows the result of multilevel modelling for the flow variable, which was the rating that showed most systematic variation. Two interesting points emerge. First, order of presentation had a major effect on rated flow. Second, a personality variable, psychoticism, had a similarly strong effect on the ratings. These are standard psychological variables. Again, it is important for system designers to understand how strong their effects are.

Conclusion

The core question for this experiment was whether the output mechanisms developed by SEMAINE affect the quality of interaction with the characters. It is clear from the ratings of flow, and still more so from the engagement traces, that they do. However, the conclusion needs to be qualified in various ways. Different types of person react differently, and reaction is dependent on what has gone before.

More fundamentally, the effect of the output mechanisms in this system seems to be character-specific. That in itself is not surprising, but there is an uncomfortable point associated with it. The characters that achieve strong engagement are negative – Spike and Obadiah. Commercially, the ability to create convincing positive characters would seem much more likely to be attractive. Hence the target of the next experiment was to establish whether manipulating output characteristics could promote positive engagement.

5.4 Experiment 5

Experiment 3 used an updated version of the SEMAINE system. A number of bugs in the previous versions were identified and addressed. More importantly the dialogue model was changed in the light of experiment 3. In particular, adjustments were made to the dialogue behaviours of Poppy and

Prudence which meant that the frequency with which they asked questions rose, and the frequency of encouragement utterances reduced.

Figure 5 shows that once again, ratings of flow are better with the full system, with $t(119) = 3.58$, $p < 0.001$. Individually, the differences are significant for Obadiah and – gratifyingly – Poppy.

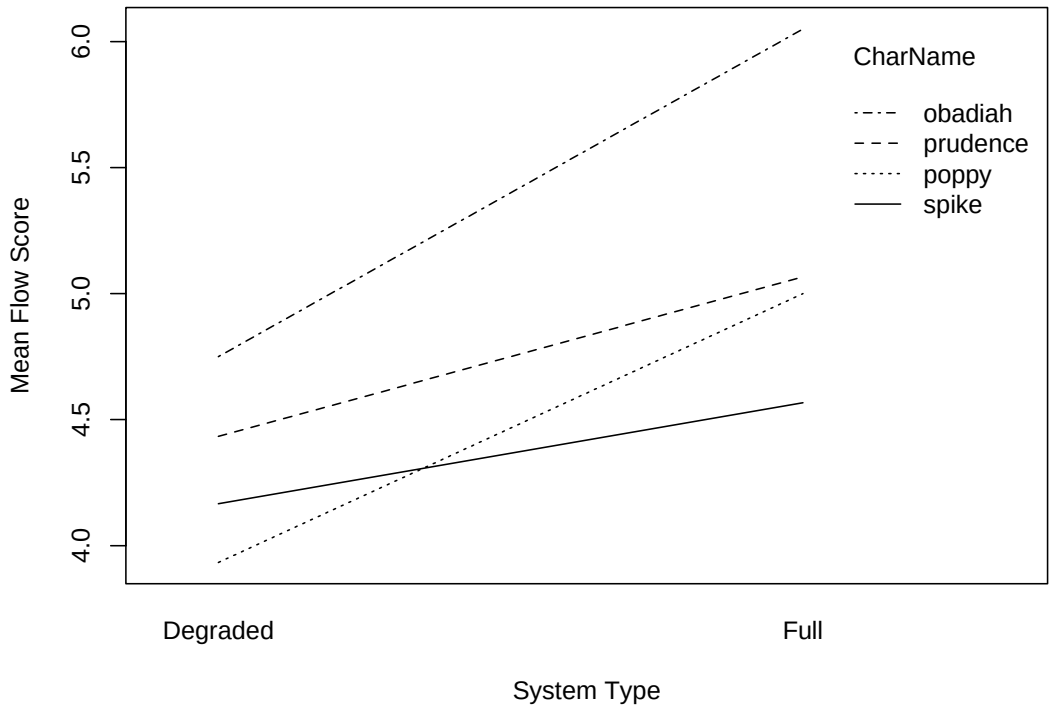


Figure 5

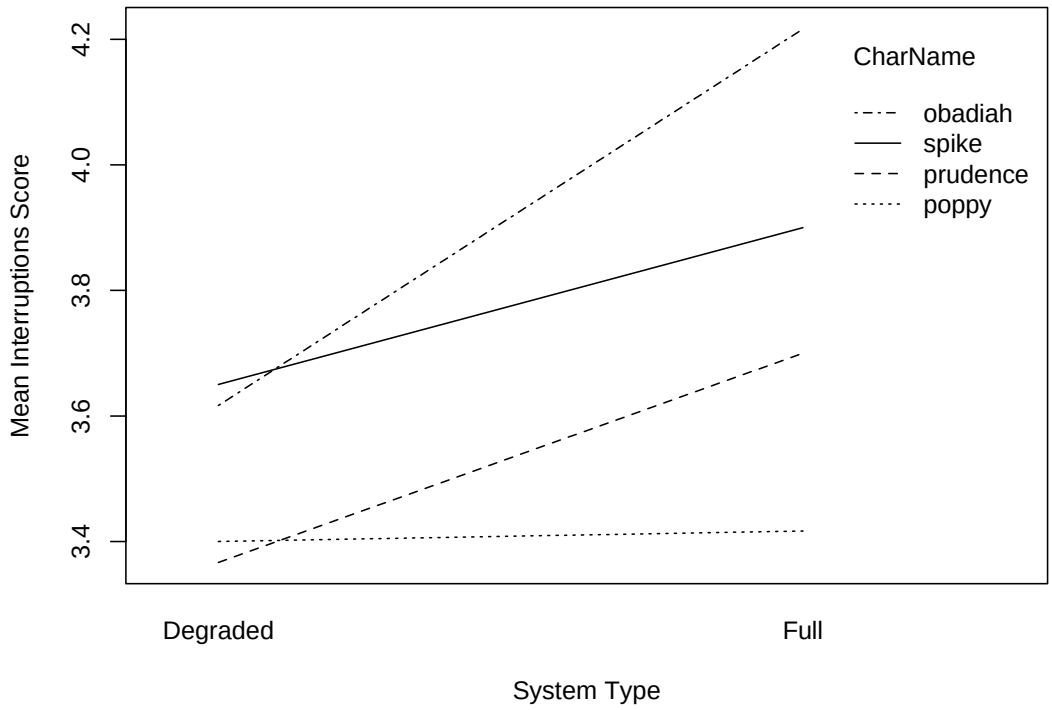


Figure 6

Also for the first time, ratings of engagement were higher for the full system, with $t(119) = 3.85$, $p < 0.001$ (Fig 7). As with competence, the differences were higher for Obadiah and Poppy individually.

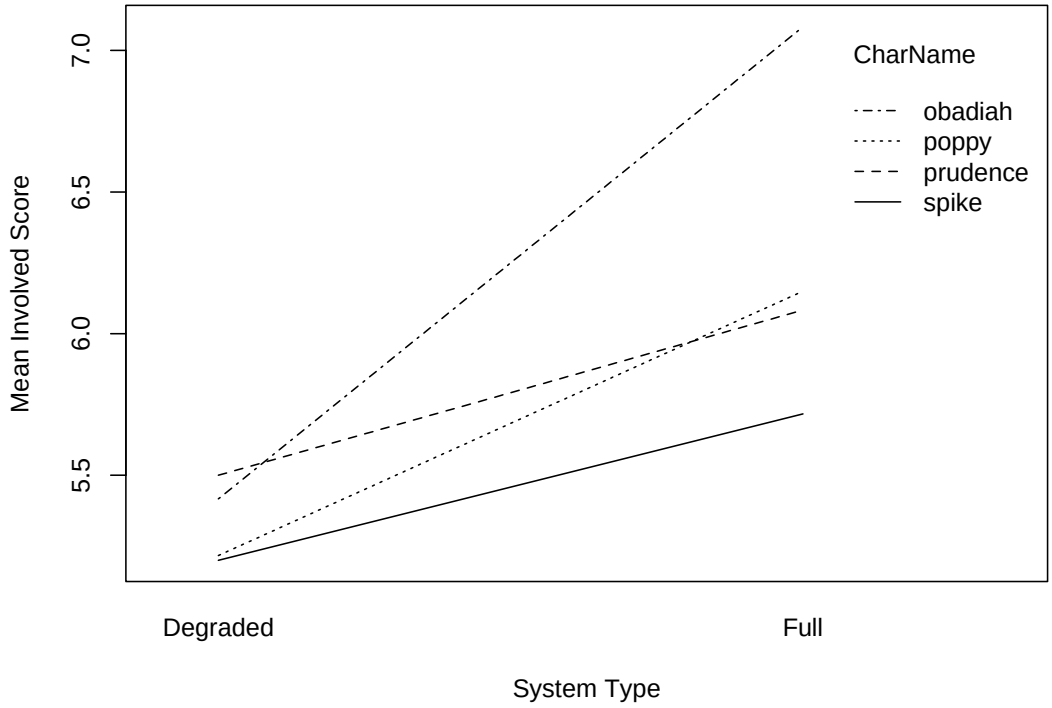


Figure 7

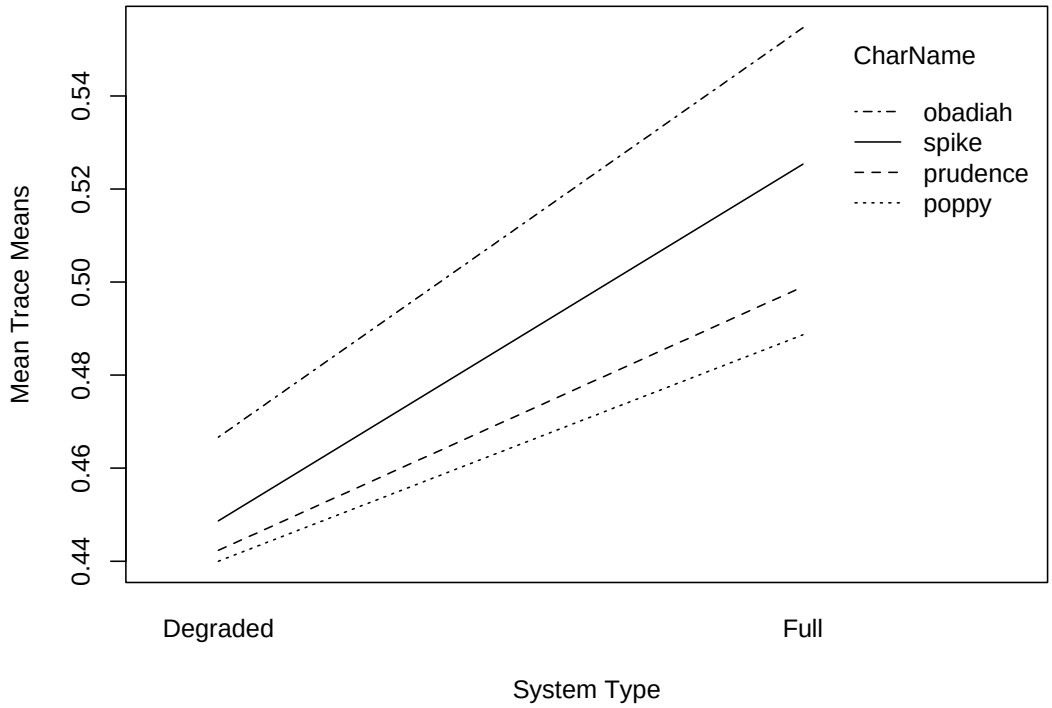


Figure 8

Traced engagement, shown in figure 8, was significantly higher overall, with $t(119) = 5.73$, $p < 0.001$, and the differences were significant for all of the individual comparisons except Poppy, where the comparison was just short of significance.

Differences in the pattern of traces were less pronounced, but once again, the systems differed in the frequency of falls in rated engagement, with $F(3,229) = 4.34$, $p=0.038$ (Figure 9).

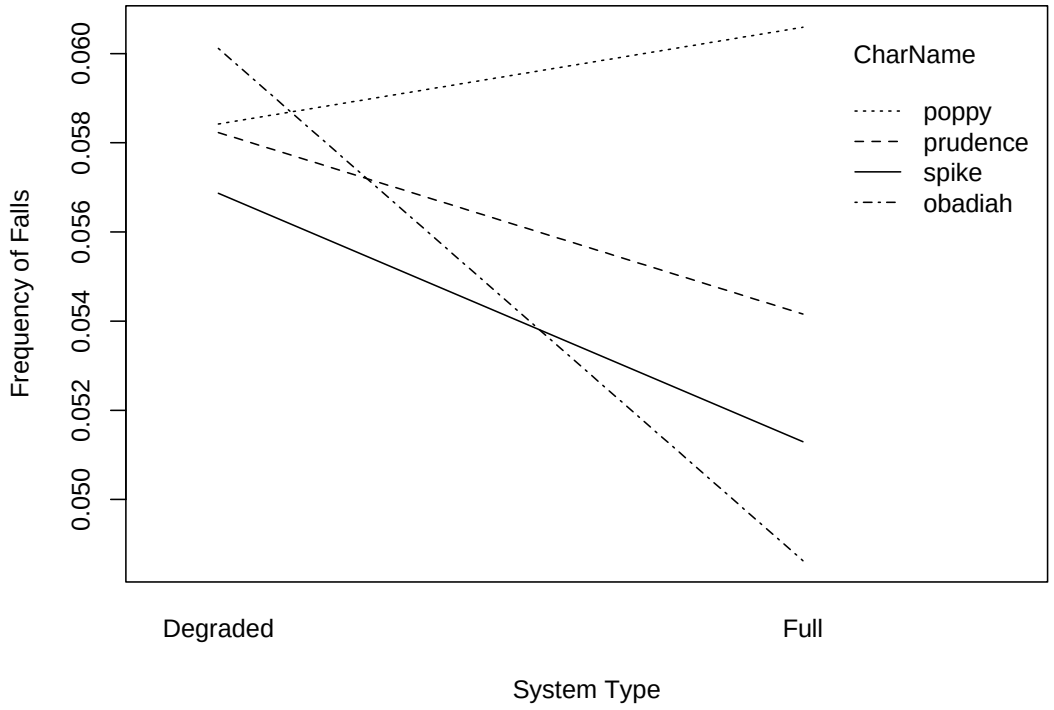


Figure 9

Table 6 shows that as before, self report and trace measures of engagement were very close, and confirms that can reasonably be regarded as measures of the same thing.

	Degraded				Full			
	Poppy	Spike	Obadiah	Prudence	Poppy	Spike	Obadiah	Prudence
self-rating	5.22	5.77	5.42	5.5	6.15	6.4	7.08	6.08
trace	0.53	0.55	0.56	0.52	0.57	0.64	0.63	0.56

Table 6

Analysis of order and personality effects revealed a pattern that appears to run through the data, but that had not been noticed before. Responses to Spike appeared to be much more negative when he was the first SAL character encountered. Table 7 shows the effect in terms of flow ratings. By chance, Spike happened to be selected relatively often in first position in experiment 5, and that probably plays a part in the relatively overall ratings for Spike in this experiment. More important practically is the striking illustration of the pattern noted in experiment 4, that responses to individ-

ual characters become more positive as the users become more familiar with the scenario. The other effects of personality and order noted in experiment 4 were not replicated here, quite possibly because the 'Spike first effect' obscured them.

Experiment	Spike First (Mean)	Spike Other Position (Mean)	Spike First Frequency	N
3	1.75	5.57	2	15
4	3.17	5.43	6	30
5	1.6	4.92	10	30
Total	2.14	5.27	-	75

Table 7 The 'Spike First' effect on flow

Users had the 'yuk button', and from observation they used it, but the routines designed to record the data from it as part of an overall log of the interaction did not function.

Simple affect measures were not significant in the previous study, but here, 29 out of 30 users expressed a preference for one of the two systems, and 24 of them preferred the full system ($p<0.002$ on a sign test).

6 Overview

The last experiment gives a clear demonstration of the core point that the SEMAINE system is preferred, on a range of measures, to a system with lower non-verbal skills. However, the picture that emerges is far more complex than that, and far more interesting.

Measure	speaker	Semiautomatic 2		Experiment 3		Experiment 4		Experiment 5	
		Degrad	Full	Degrad	Full	Degrad	Full	Degrad	Full
flow	Poppy	5.33	5.83	4.63	4.57	5.12	5	3.93	5
	Spike	3.6	6.25	5.47	5.17	4.65	5.75	4.17	4.57
	Obadiah	5	4	5.77	4.77	4.77	5.98	4.75	6.05
	Prudence	5.6	3.5	6.2	4.5	4.38	4.33	4.43	5.07
	Overall	4.88	4.90	5.52	4.75	4.73	5.27	4.32	5.17
correctness	Poppy	3.33	4.33	3.20	3.80	3.73	3.82	3.40	3.42
	Spike	2.40	3.75	3.40	3.80	3.42	3.90	3.65	3.90
	Obadiah	3.60	4.00	3.73	3.73	3.75	3.88	3.62	4.22
	Prudence	4.00	3.00	3.79	4.07	3.78	3.66	3.37	3.70
	Overall	3.33	3.77	3.53	3.85	3.67	3.82	3.51	3.81
engagement	Poppy	4.67	6.17	5.47	5.86	5.8	5.82	5.22	6.15
	Spike	3.8	6.25	6.03	7.07	5.77	6.4	5.77	6.4
	Obadiah	5.6	4.75	6.73	6	5.95	6.35	5.42	7.08
	Prudence	6.4	3.5	6.8	6.27	5.23	5.52	5.5	6.08
	Overall	5.12	5.17	6.26	6.30	5.69	6.02	5.48	6.43

Table 8

Table 8 is a useful starting point. It uses a colour code to bring out key patterns. In each row, the highest rating is coloured green; the next highest yellow; the lowest red; and the next lowest orange.

It stands out immediately that the worst ratings are concentrated in two columns, the first and the second last. The low ratings in the first column are informative because the operator is a human, meaning that the higher ratings in the later columns show the automatic system outperforming a human operator – albeit one whose feedback from the user involves only vision. The ratings in the second last are similar to the experiment 4 degraded condition, as they should be, because the degraded systems are the same. The fact that both are low underlines the point that the various output functionalities of SAL make a difference to the system's reception, because it is the output functionality that is degraded in these systems. The apparent difference between them underscores a point that the statistics have made at various stages, which is that responses to the agents are context-dependent. It is very likely that the degraded system in experiment 5 receives lower scores because it is contrasted with a more competent full system.

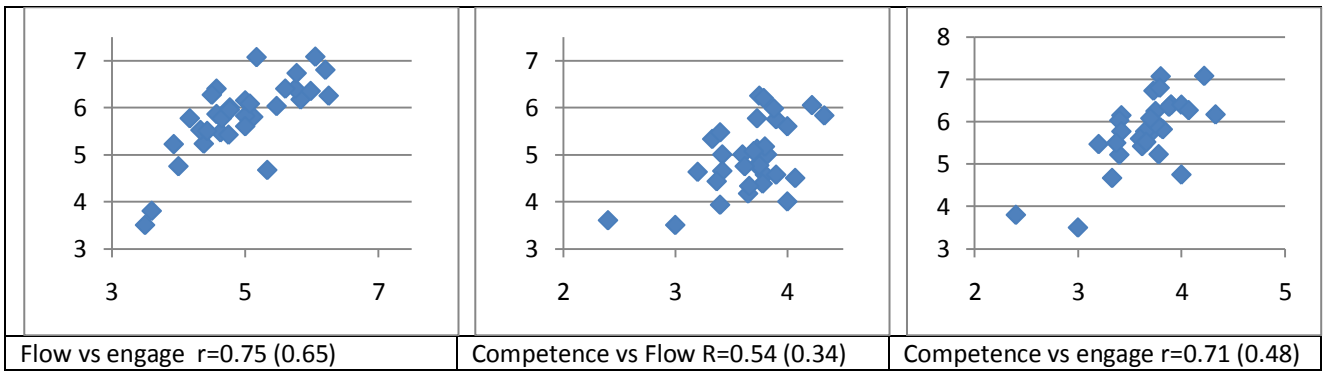
Two particular cases cut across these general trends. The best ratings for Poppy come from experiment 2, where a human being with full feedback is choosing what the character says. The most striking point is that that only happens with one character. That is a considerable endorsement of the system. However, it is also interesting, and important, to ask why it happens with that particular character. It may be that a negative character can be convincing without appearing to register anything that the user is saying, but a positive character cannot.

A similar issue arises with Prudence. Her best ratings come in degraded systems, with experiment 3 best overall, and experiment 2 (with semiautomatic SAL) following. The natural explanation for the high ratings on experiment 3 is that Prudence is most acceptable when her habitual reasonableness is broken by moments of irrationality, when she makes statements that are quite at odds with the user's actual feelings. There is support for that in the very low ratings that she receives when a human operator with full feedback is choosing her utterances (in experiment 2). It seems that the more Prudence acts in character, the less people like her. That has important practical implications, because Prudence-like characters – projecting competence and rationality – are the sort that it is natural to imagine in applications. That may not be as good a choice as people might automatically assume.

These particular points underscore two general issues which are extremely important for the development of affective avatars.

The first is that avatar personality plays a major part in people's likes and dislikes. It is not simply a matter of which personality people like. Competences that make one personality more acceptable make another less so, and a personality that is difficult to deal with initially may be well received by people who are familiar with the system. These points can emerge because avatar personality was included as a variable in SEMAINE. It appears to have been a good decision, because not only does avatar personality matter: the empirical evidence suggests that behaves in ways that few people would have predicted a priori. The point has already been made that the Prudence character may not be a good model for working systems. Conversely, the characters dominated by negative emotions seem to be the best received. That may simply be a quirk of the particular characters that emerge from the SAL scripts, but it deserves to be considered carefully.

The second point is that as in other areas, the relationship between evaluation and design is not straightforward. It is clearly important to ask whether the SAL characters are well enough received to be at all interesting. But given that it seems reasonably clear that they are, the obvious way to look at data from users is not as a measure of how good the system is, but as a guide to ways of making it better. Particular ideas about making the system better are considered in the 'directions' section below. More generally, though, thinking in terms of continuous improvement brings into play well-known arguments about the kind of empirical work that is needed to inform design (Kaye et al 2011 In that context, set piece experiments designed to show significant effects are probably less useful than smaller studies that can be done quickly, to gain information about a particular version of the system in order to modify it. It follows that the evaluation techniques developed here need to be thought of in relation to the design cycle as well as tools for set piece studies.



Flow vs engage r=0.75 (0.65)	Competence vs Flow R=0.54 (0.34)	Competence vs engage r=0.71 (0.48)

Figure 10

Turning to the evaluation techniques, figure 10 shows the relationships between them in terms of

the cells in Table 8. It is clear that flow and engagement are closely related, and that competence is less closely related to either. The issue is complicated by the fact that two data points are distinctly different from the rest. They correspond to Spike and Prudence in semiautomatic SAL interactions, Spike with the degraded system and Prudence with the full version. These were clearly unacceptable in a way that was quite different from other interactions, and all the measures seem to agree on that. If those extreme cases are excluded, the picture changes quite substantially. The correlations shown below each graph make the point. In each case, the first value is the correlation for all points, including the extreme cases. The second is the correlation excluding the extreme cases. That indicates that outside the extremes, flow and competence are not at all strongly correlated – that is, the 'feel' of the system cannot be deduced from its competence.

The flow question is consistently best at differentiating between systems. That is not unexpected, given the evidence in the literature that people find flow judgments natural. Table 8 makes an interesting point about flow, which is that it seems to be most effective at picking out systems which are not satisfactory. The engagement measure shows weaker relationships to system characteristics throughout. Given the pattern in figure 10, it seems quite likely that that is because the measure is less natural: people have difficulties recalling and verbalising their own states, even very shortly afterwards.

The interesting point about engagement is that it can be measured in a different way, by asking an external observer to rate apparent engagement. The relationships between external and self-ratings of engagement, which were presented earlier, suggest that they measure essentially the same thing. However, the external rater's assessment relates more directly to system characteristics than the self-ratings do. The most obvious explanation is that it is a less noisy measure.

The traces have an additional advantage, which is that they provide information on change of engagement within an encounter. It has been shown that the SEMAINE stylisation technique provides a way to access some of that information. It make sense that falls in engagement are a particularly telling measure. That clearly has potential as a design tool, since it means that system builders can home in on regions of an interaction where something has gone wrong, and try to identify what it is and how it can be improved.

A final point that the techniques bring out has implications for any kind of evaluation. The user has to be recognised as a variable in evaluations of this kind of system. That is true at two levels. Stable traits of the user, as measured by personality tests, have a substantial effect on response. It seems likely that some people will not react in this kind of situation however socially skilled the system. Transient states also have an effect. On one hand, people adjust to the unusual kind of interaction. On the other, first interactions in particular can dispose people positively or negatively towards the system. One would expect formal evaluations to balance for effects like these, but they also need to be taken into account in design-oriented studies. It would not be reasonable to assume that the reactions of a particular individual, perhaps with extensive experience of the system, could stand for the population at large.

These issues are food for a great deal of thought. However, it is important not to lose sight of the most basic finding of the evaluation. It is simply that the SEMAINE system clearly can engage people in sustained, emotionally coloured conversations. A number of participants commented after the interaction how extraordinary they found that, expressing disbelief that they had just spent tens of minutes engrossed in a conversation with a computer. The system is deeply limited, but it demonstrates that emotion-rich, spoken conversations with avatars are a real possibility rather than science fiction.

7 Directions

One of the outcomes of evaluation is to highlight natural lines of development. The specifics are often related to a combination of sources, the formal evaluations reported above and impressionistic interpretation of the responses that appear in recordings.

The SAL scenario was explicitly designed to pave the way for exploration of spoken interactions between a human and an avatar. To do that, SEMAINE created a special situation where limitations that could not be overcome – particularly limitations related to language – would not be fatal. Seeing the resulting interactions, it becomes clear that the scenario highlights skills which are not restricted to the particular, artificial scenario, but which it provides a way to develop. Five in particular stand out: knowing when to come in; taking the other party's state as a topic of conversation; 'drawing out' the other party to express views and feelings about a topic; understanding what makes an acceptable sequence of utterances (particularly, but not exclusively, in the course of exchanges that are drawing the other party out or talking about his/her state); and showing emotional flexibility. The point about these skills is not just that one can imagine making progress on them, but that they are potentially useful to a functional system.

Observing SAL interactions points up the way that these depend on various lower level developments. For example, it is clearly important to find ways of recognising states that are not simply emotional. One that stands out is thinking or puzzling about something. It is obviously related to the issue of taking another party's state as a topic of conversation: the obvious question is something like "what are you thinking about?" Less obviously, it is related to the problem of knowing when to come in. One of the mistakes that SAL agents typically make is to intervene when it is obvious to a human that the user is thinking about what to say. Ignoring the signs that convey that is conversationally incompetent.

Showing emotional flexibility is a very different example. It is an obvious extension of the way the avatars operate. They change their utterances according to the user's state, but their own emotional orientation does not move at all. It is very obvious from the interactions that after a while, Spike's unremitting aggression, or Prudence's unremitting reasonableness, jar on users. Their emotional balance should shift at a deeper level in response to some of the signals that users give. It is not difficult to imagine mechanisms that would simulate that kind of shift at least crudely.

All of this reflects the general point that once systems like SAL exist, a great variety of other issues can be understood in terms of tuning the systems. Certainly some kinds of tuning are trivial, and of no wider relevance; but there is no shortage of challenges that are of general interest.

To make the most of that kind of strategy, the system needs to be set up in a way that allows relatively self-contained adjustments to be made, and their effects evaluated. One of the problems that evaluation exposed was that the original conception of SAL did not envisage the kind of progressive adjustment that has been considered in the last two sections. The problem is not insuperable. Because the system is fundamentally modular, there are ways of changing one ability at a time, so that the effects can be observed. Like many other things, the importance of building in that kind of flexibility only becomes clear when a working system exists, and it becomes possible to see the kind of process that is needed to transform it from an engaging prototype into a functional conversation partner.

8 References

R Cowie Perception of emotion: towards a realistic understanding of the task. Phil. Trans. R. Soc. B (2009) 364, 3515-3525

M. Csikszentmihalyi and I. S. Csikszentmihalyi, Beyond boredom and anxiety. San Francisco, USA: Jossey-Bass, 1975.

M. Edwardson, "Measuring consumer emotions in service encounters: an exploratory analysis," Australasian Journal of Market Research, vol. 6, no. 2, p. 34–48, 1998.

K. Isbister, K. Hook, M. Sharp, and J. Laaksolahti, "The sensual evaluation instrument: developing an affective evaluation tool," in Proceedings of the SIGCHI conference on Human Factors in computing systems. Montr´eal, Qu´ebec, Canada: ACM, 2006, pp. 1163–1172. [Online]. Available: http://portal.acm.org/citation.cfm?id=1124772.1124946

ISO 1998 I. I. O. for Standardization, "Ergonomic requirements for office work with visual display terminals (VDTs) - part 11: Guidance on usability," International Standards Organisation, ISO Standard ISO 9241-11, 1998.

J. J.Kaye, J. Laaksolahti, K. Höök, and Ka. Isbister (2011) The Design and Evaluation Process In Petta, Pelachaud & Cowie (ed) *Emotion-Oriented Systems: The Humaine Handbook* Berlin: Springer pp 637-652.

Pennebaker, J. W., & Chung, C. K. (2007) Expressive Writing, emotional upheavals, and Health In H. S. Friedman & R.C Silver (ed) (Ed.), *Foundations of health psychology*. New York, NY: Oxford University Press pp. 263-284

M. V. Sanchez-Vives and M. Slater, "From presence to consciousness through virtual reality," Nature Reviews Neuroscience, vol. 6, no. 4, pp. 332–339, 2005. [Online]. Available: http://dx.doi.org/10.1038/nrn1651

M. Schroeder & G. McKeown Considering social and emotional artificial intelligence. Proc AISB symposium "Towards a comprehensive Turing Test". Leicester, 2010.

S. Westerman, P. Gardner, and E. Sutherland, "Usability testing Emotion-Oriented computing systems: Psychometric assessment," HUMAINE deliverable D9f, 2006. [Online]. Available: http://emotion-research.net/projects/humaine/ deliverables/D9f%20Psychometrics%20-%20Final%20-% 20with%20updated%20references.pdf