Evaluating a Spoken Dialogue System for recording clinical observations during an endoscopic examination

D. J. BARKER†*, P. VAN SCHAIK‡, D. S. SIMPSON† and W. A. CORBETT§

†School of Computing and Mathematics, University of Teesside, Middlesborough, Tees Valley, TS1 3BA, UK
‡School of Social Sciences and Law, University of Teesside, Middlesborough, Tees Valley, TS1 3BA, UK
§Endoscopy Centre, James Cook University Hospital, Middlesborough, Cleveland, UK

(Received June 2002)

Abstract. Paper, keyboard or mouse-driven systems may not be suitable for data capture because of the hands-busy constraint imposed by an endoscopic examination. A Spoken Dialogue System (SDS) has a number of advantages when compared to keyboard and mouse-driven input modalities, particularly with respect to hands free and eyes-free control of a system. However, any emerging technology will never deliver improved organizational effectiveness if it is not accepted and used [1]. The Technology Acceptance Model (TAM) provides a framework that helps explain the determinants of computer acceptance [2]. This study, through the application of TAM, demonstrates a high level of user acceptance with clinicians wanting to use spoken dialogue technology for recording clinical observations during an endoscopic examination. Clinicians would also prefer to use a SDS for recording endoscopy rather than use a paper-based or keyboard and mouse-driven system. Using a clinical narrative during an endoscopic examination was also perceived to be a natural way to record findings. Relationships between basic TAM variables were confirmed and relationships between quality of dialogue measures and TAM variables were established.

Keywords: Spoken Dialogue System; Endoscopy; Evaluation; Technology Acceptance

1. Introduction

Endoscopy is a general term which describes the internal examination of the upper or lower gastrointestinal (GI) tract. Immediately after a patient encounter, clinicians at the James Cook University Hospital (JCUH) use a paper form to document any observations made or procedures undertaken during the examination. The information recorded on the paper form is then transferred to a relational database, which is held within the department. The task of transferring data from the paper form to the relational database is not undertaken by the clinician performing the endoscopy, but by data entry personnel. The delay between the original endoscopic examination and the transfer process may, in extreme cases, be many months and only serves to hinder the care process, because timely patient information is not available to clinicians over the department’s computer network at the time and place it is needed. In addition, a limited knowledge of the clinical domain when transferring information must always raise questions about the
integrity of any data that is stored on the system. It is also unlikely that data entry personnel represent an efficient use of finite healthcare resources.

One possible solution to the problem would give clinical access to the computer system at the end of an examination, rather than use data entry personnel to input information retrospectively. Grasso [3] observes that end users are likely to abandon an application if it does not improve or possibly interferes with their work and this is a view supported by clinicians from the department. Although they acknowledge the value of recording data onto the computerized system, they feel that the current keyboard-driven text-based application used by data entry personnel is difficult to learn and difficult to use. For clinicians, text-based or even mouse-driven data entry require the use of the hands. Therefore, these systems can only be used to record information after a hands-busy clinical procedure such as endoscopy. However, video recordings taken during endoscopic examinations have clearly highlighted the danger of not being able to record information at source [4]. For example, biopsies taken during one examination had not been recorded on the paper form after the examination had finished. In short; paper, keyboard or mouse-driven systems are not suitable for data capture because of the hands-busy constraint imposed by an endoscopic examination.

Observational and procedural narratives are a natural way for clinicians to communicate clinical information. An example of an observational narrative would be ‘large ulcer found located 10 cm from the cardio-oesophageal junction’. An example of a procedural narrative would be ‘multiple biopsies taken at site of ulcer’. A speech interface has a number of advantages when compared to keyboard and mouse-driven input modalities, particularly with respect to hands-free and eyes-free control of a system. However, the complex hidden structure of a natural language means that speech recognition as a single input modality is unable to capture the information carried in a sentence that depends on the relationship between words. If a speech interface is to be used for recording endoscopy, identifying these relationships is essential for two main reasons. Firstly, if a relational database is to be used to store patient data, the speech input must be broken up into a structure that can be mapped onto the target database schema. For example, a finding needs to be stored in a finding field and its position in a position field. Secondly, spoken language interaction between humans depends on the ability of each participant to understand each other in order to maintain an appropriate dialogue. For example, in an endoscopy context the phrase ‘malignant ulcer seen’ is incomplete, because it requires a positional qualifier. An appropriate clinical response to this statement would be to ask where was the malignant ulcer seen. A system that enables humans to interact with a computer application using a natural spoken language is referred to in the literature as a Spoken Dialogue System (SDS). A SDS is comprised of four main components: speech recognition, natural language understanding, dialogue management and response output [5–7].

This research supports the view by Berwick [8], that every system is designed to achieve the results it achieves, with performance being a product of design and not just effort. A SDS is by design more suitable for data capture during endoscopy when compared to paper, keyboard or mouse-driven systems because (a) it would allow clinicians to record information at source during a hands-busy examination and (b) it would enable clinicians to easily record observations using a natural clinical language. Spoken dialogue technology can address the limitations imposed by keyboard and mouse-driven interface metaphors, but available systems are currently experimental [9].
However, any emerging technology will never deliver improved organizational effectiveness if it is not accepted and used [1]. The Technology Acceptance Model (TAM) provides a framework that helps explain the determinants of computer acceptance, which is capable of explaining user behaviour across a range of computer technologies [2]. TAM (see figure 1) also enables the impact of external factors on internal beliefs, attitudes and intentions to be examined. Two particular beliefs, perceived usefulness and perceived ease of use, have been shown to be of primary importance for usage behaviours.

According to TAM, perceived usefulness (PUSF) and perceived ease of use (PEOU) are determinants towards usage intentions and actual technology use [2]. Perceived usefulness is defined as ‘the degree to which a person believes that using a particular system would enhance his or her job performance’ while perceived ease of use is defined as ‘the degree to which a person believes that using a particular system would be free of effort’ [2]. TAM suggests that actual usage is determined by a behavioural intention to use a system, which is determined by perceived usefulness and perceived ease of use. Perceived usefulness is influenced by perceived ease of use and both of these are influenced by external variables, including product characteristics. According to TAM, the easier a technology is to use and the more useful it is perceived to be, results in a more positive intention towards using the technology.

More recent research has extended TAM with additional determinants of PUSF and PEOU [10]. However, as the current study was the first of its kind to investigate the acceptance of a SDS, it was decided to establish the applicability of TAM to this type of system by focusing on basic TAM variables and their relationships.

The evaluation study does not aim to determine clinical acceptance with respect to reports generated by the SDS. This is because in the current prototype system, no reports can be generated. In addition, no structured data can be saved by the SDS to a clinical database. The aims of the evaluation study are to (a) determine whether or not clinicians would want use a SDS during an endoscopic examination by application of the TAM framework, (b) determine if clinicians would prefer to use a SDS for recording endoscopy, rather than paper-based or keyboard and mouse-driven systems, (c) examine the impact of the external variables that influence TAM and (d) examine additional measures of perceived system quality.

Figure 1. Technology acceptance model [1]
2. Methodology

2.1. Materials

Two sessions were used to collect user and system data: enrolment and evaluation. The enrolment session questionnaire collected demographic data and also user experience with computers and speech recognition. Enrolment was performed on a different day from the evaluation to avoid stressing the vocal tract, as the enrolment process involved reading a pre-prepared script to enable the recognizer to build a unique voice model for each participant. The standard script was provided by IBM for use with their ViaVoice product and contained 88 sentences, which take about 20–35 min for a user to read. Another 5 min was then required to process the captured data and prepare the voice model.

An evaluation questionnaire (see appendix) measured technology acceptance in terms of perceived ease of use (questionnaire items 10, 11, 12 and 13, based on [2]), perceived usefulness (questionnaire items 14, 15, 16 and 17, based on [2]) and intention to use (questionnaire items 18, 19, 20 and 21, based on [11, 12]). The questionnaire also measured perceived system performance (questionnaire items 1, 2 and 3), perceived speed of dialogue in terms of user turns (questionnaire items 4 and 5) and perceived system response time (questionnaire items 7 and 8). Additional measures of perceived system quality were also included (questionnaire items 6, 9, 22, 23, 24 and 25). All items used a seven point Likert scale, with end points of strongly disagree (−3) to strongly agree (3) and a neutral answer of 0.

The evaluation required a number of scenarios or test cases, in order for clinicians to present an observation to the system. The test cases were presented in a booklet, with each case presented on a separate page. Each test case consisted of a pictorial representation of a finding placed in the upper gastrointestinal tract. A graphical representation was chosen over a text-based scenario, in order to limit lexical and grammatical influences. However, certain clinical attributes were described using text (for example, benign or bleeding), as this information was not clear from the images used.

The choice of evaluation set was based on a statistical analysis of approximately 18,550 endoscopies taken from July 1989 to July 1996. The current system records 65 different items for the upper gastrointestinal tract and occur with different frequencies. For example, oesophagitis represents some 13.62% and a hiatus hernia 11.75% of all observations made. Normal stomach is 28.03%. If the test set focused on these three items, it would represent 53.40% of all observations made. However, the overall coverage would only be 4.60% (that is, three items out of a possible 65). It was found that the set of ulcer findings only represents 11.52% of all observations made, but their coverage is 40% (that is, 26 items out of 65). A large coverage is preferred for evaluating the prototype Spoken Dialogue System (SDS), as it is likely to induce more variation in the language used by clinicians, rather than a small coverage group. Because ulcers are the largest set of findings recorded by the current system, they were chosen in preference to other findings for the evaluation text cases.

A test set of 26 ulcer findings was used for the evaluation. Statistically, 3.07% of recorded ulcers are found in the oesophagus, with 53.85% and 38.46% being located in the stomach and duodenum respectively. For the evaluation set, this represents two, 14 and 10 items for the oesophagus, stomach and duodenum respectively. Three additional cases were used for training users on the system. Unlike the
evaluation cases, each training case had four examples of the types of language that could be used in order to record a finding.

Although the evaluation test set was limited to ulcer findings, it should be noted that the SDS lexicon, language models and domain model were not constrained for the study. That is, the evaluation system modelled language and domain understanding for the upper gastrointestinal tract with 65 findings. However, no therapeutic terms (for example, taking a biopsy) were modelled during the term of this research project.

2.2. Apparatus

The evaluation system comprised of a Personal Computer (PC), with an 800 MHz processor, 256MB RAM and 12.2 GB hard drive running Windows NT with service pack 6. A Creative Labs Sound Blaster (SB) Live sound card was used to capture audio, driven by a balanced line receiver. The receiver was connected to a balanced line transmission unit with built in Automatic Gain Control (AGC), via a studio grade cable. A noise cancelling headset microphone fed the transceiver, with the AGC compression slope set to approximately 4:1. With the exception of the speech recognizer engine (IBM ViaVoice Millennium Edition, version 7), all of the prototype SDS software was developed specifically for this research project.

2.3. Subjects

Ten endoscopists were involved in the study, all of them male. The age range of the participants was 20–29 (2), 30–39 (3), 40–49 (3) and 50–59 (2). Six of the clinicians were native English speakers, the remainder having mixed native language backgrounds, which comprised of Persian (1), Tamil (1), Greek (1) and Urdu (1). All participants reported having experience with using computers, which ranged from 2–21 years. Four of the participants reported having used speech recognition technology, ranging from 6 months to 4 years. All of the subjects were volunteers.

2.4. Procedure

2.4.1. The enrolment session. Each subject was required to enrol on the system in order to generate a personal voice model for use by the speech recognizer during the evaluation session. Before reading the script, the headset microphone was correctly positioned and a sound level taken of the subject speaking and adjusted accordingly. At the end of the enrolment session, the enrolment questionnaire was completed by the subject while the computer processed the voice model data. The time taken for the enrolment session varied between each participant, ranging from 20 to about 35 min. The variability in time taken to complete a script was due to (a) the speed in which each participant would read the enrolment script and (b) how often a subject was asked to repeat a phrase by the script control software.

2.4.2. The evaluation session. The evaluation session commenced by explaining the function of the system and also the purpose of the evaluation. Each participant was advised that the SDS had been designed to record findings during an endoscopic examination and that it was not designed to be a medical dictation system. That is, the system would not understand the type of language used when dictating a discharge letter and phrases should generally be kept as short as possible. Clini-
cians were also informed that the purpose of the evaluation was to test the usefulness of the system and not how well they performed during endoscopy.

The next phase of the evaluation session involved giving each participant hands on experience in using the system to record a finding. This was achieved by using three training cases, which were identical in format to the 26 evaluation cases. Each training case was supplemented with four example utterances that the system would understand, giving 12 phrases in total. Clinicians were asked to record an observation onto the system using the 12 sample phrases. During the training period, subjects were encouraged to freely ask questions relating to the use and operation of the system.

The final phase of the evaluation involved recording observations onto the system using the 26 test cases, during which time no questions were allowed. Participants were advised that they were free to use their own clinical judgement to interpret each test case. If the system did not understand what was said during the evaluation, or it did not record an observation correctly, then one of three options were available; (a) re-present the observation, that is, say it again, (b) rephrase the question and say it again, or (c) ignore the problem and turn the page to the next test case. At the end of the evaluation session, the subject was asked to complete the evaluation questionnaire.

3. Results

The findings of the evaluation study are described mostly using the framework of the Technology Acceptance Model [2]. Firstly, basic TAM variables are analysed. Secondly, an analysis of variables influencing basic TAM variables is conducted. Third, additional measures of perceived system quality are analysed.

3.1. Basic TAM variables

Reliability was assessed using Cronbach’s alpha for the four items (refer to appendix for a complete description of each question) used to construct perceived usefulness (PUSF; questionnaire items 14, 15, 16, and 17), perceived ease of use (PEOU; questionnaire items 10, 11, 12, and 13) and intention to use (INTENT; questionnaire items 18, 19, 20, and 21). The reliability coefficient for each construct was 0.96, 0.91 and 0.97 respectively. Descriptive statistics for each construct indicated that subjects perceived the SDS as easy to use (Mean = 1.15, SD = 1.34), useful (Mean = 1.58, SD = 1.49) and that they intended to use the system if it were to become available (Mean = 2.20, SD = 1.05). These findings were confirmed by t-tests: $t(9) = 2.70$, $p < 0.05$ for PEOU and $t(9) = 3.35$, $t(9) = 6.60$ for PUSF and INTENT respectively, both $p < 0.01$.

Regression analysis was performed on PUSF and INTENT, shown in table 1. The analysis showed that perceived ease of use was a significant predictor of perceived usefulness and that perceived usefulness was a significant predictor of intention to use. A further multiple regression was performed on INTENT, using both PUSF and PEOU as regressors. PEOU was found not to be an additional significant predictor of INTENT, with $R^2_{change} = 0.002$.

3.2. Variables influencing basic TAM variables

Reliability was assessed using Cronbach’s alpha for the items (refer to appendix for a complete description of each question) used to construct perceived system
performance (SYSPERF), perceived speed of dialogue in terms of user turns (SPEEDDLG) and perceived response time (RESPTIME). Three items were used to construct SYSPERF (questionnaire items 1, 2, and 3) and two items for SPEEDDLG (questionnaire items 4 and 5) and RESPTIME (questionnaire items 7 and 8). The reliability coefficient for SYSPERF, SPEEDDLG and RESPTIME was 0.95, 0.90 and 0.96 respectively. Descriptive statistics for each construct indicated that perception of system performance (Mean = 0.93, SD = 1.20), speed of dialogue (Mean = 1.25, SD = 1.38) and system response time (Mean = 1.35, SD = 0.75) were positive. These findings were confirmed by \( t \)-tests: \( t(9) = 2.45, p < 0.05 \) for SYSPERF and \( t(9) = 5.71, p < 0.01 \) for RESPTIME.

A multiple regression was performed on PUSF, using both PEOU and SYSPERF as regressors. SYSPERF was found not to be an additional significant predictor of PUSF, with \( R^2_{\text{change}} = 0.04 \). However, a significant positive correlation between SYSPERF and PUSF was observed, with \( r = 0.72, p < 0.05 \). Further regression analysis was performed on PEOU, shown in table 2. The analysis showed that the perceived speed of dialogue in terms of user turns was a significant predictor of perceived ease of use. Perceived system response time was found not to be significantly correlated with perceived ease of use, with \( r = 0.17, p > 0.05 \).

A number of system performance measures were analysed in relation to the model. These included the total number of dialogue failures per session (DLGFAIL; Mean = 0.60, SD = 0.97), the mean number of user turns per session (AVTURNS; Mean = 1.53, SD = 0.26) and the maximum dialogue turn per session (MAXTURNS; Mean = 3.90, SD = 1.60). Regression analysis showed that MAXTURNS was a significant predictor of perceived speed of dialogue, shown in table 2.

The mean number of user turns per session was found not to be significantly correlated with perceived speed of dialogue, with \( r = -0.58, p > 0.05 \). A further

| Table 1. Regressions for PUSF and INTENT |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Dependant variable | \( R^2 \) | \( F \)-statistic | Independent variable | \( b \) | \( \beta \) | \( t \)-statistic |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| PUSF            | 0.50            | 8.10*           | PEOU            | 0.78           | 0.71           | 2.85*           |
|                 |                 |                 | Constant        | 0.67           |               |               |
| INTENT          | 0.61            | 12.74**         | PUSF            | 0.56           | 0.78           | 3.57**         |
|                 |                 |                 | Constant        | 1.33           |               |               |

\( n = 10 \)

\* \( p < 0.05 \)

\** \( p < 0.01 \)

| Table 2. Regression for PEOU and SPEEDDLG |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Dependant variable | \( R^2 \) | \( F \)-statistic | Independent variable | \( b \) | \( \beta \) | \( t \)-statistic |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| PEOU            | 0.58            | 11.20*          | SPEEDDLG        | 0.75           | 0.76           | 3.35*           |
|                 |                 |                 | Constant        | 0.22           |               |               |
| SPEEDDLG        | 0.52            | 8.59*           | MAXTURNS        | -0.62          | -0.72          | -2.93*          |
|                 |                 |                 | Constant        | 3.68           |               |               |

\( n = 10 \)

\* \( p < 0.05 \)
multiple regression was performed on PEOU, using both MAXTURNS and SPEEDDLG as regressors. MAXTURNS was found not to be an additional significant predictor, with $R^2_{change} = 0.09$. However, a significant correlation between MAXTURNS and PEOU was observed, with $r = -0.76$, $p < 0.05$. Regression analysis also showed that the total number of dialogue failures per session was not a significant predictor of perceived system performance, with $F(1,8) = 1.31$, $p = 0.29$ for $R^2 = 0.14$. The TAM for the prototype SDS is shown in figure 2.

3.3. Additional measures of perceived system quality

Additional questions to measure perceived system quality indicated that subjects found: using a clinical narrative was a natural way to record endoscopic findings; the SDS enabled them to work in a way that they like to work; did not become impatient when using the SDS; prefer to use the SDS rather than a paper form at the end of an examination; prefer to use the SDS rather than a keyboard and mouse at the end of an examination. No agreement between subjects was found as to whether the SDS requested appropriate information before recording a finding. These results were confirmed using a Wilcoxon signed rank test, shown in table 3, tested against a neutral score of zero.

4. Discussion

Results showed high levels of technology acceptance, with a clear indication that clinicians would use the SDS were it to become available for use during endoscopic examinations. Clinicians would prefer to use the SDS to record endoscopic observations, rather than the current method of having to record information at the end of an examination using a paper form. In addition, clinicians would prefer to use the SDS to record endoscopic observations, rather than having to use a keyboard and mouse. The process of using a clinical narrative was also found to be a natural way to record endoscopic findings and using the SDS enabled clinicians to work in a way that they like to work.

The relationships between the TAM components confirmed findings by previous research [1]. These were; (a) a positive relationship between perceived ease
of use and perceived usefulness, and (b) a positive relationship between perceived usefulness and intention to use. Evaluations of TAM have shown a much stronger relationship between perceived usefulness and intention to use, when compared to the small but significant relationship between perceived ease of use and intention to use. However, in a study to evaluate clinical acceptance of a low cost system for postural assessment [11], the relationship between perceived ease of use and intention to use was found to be not significant.

It was suggested that this difference in outcome might be due to respondents not actually having hands-on experience with the system they were rating, but the SDS evaluation also found that perceived ease of use was not a significant additional predictor of intention to use. An alternative explanation is that this study, like the one by van Schaik et al., had a more specialized subject population which consisted of endoscopists and physiotherapists respectively. This is in contrast to other studies, having evaluated TAM against Information Technologies (IT) with a more general population, for example, word processors, mail applications and web browsers. Specialists in their work may therefore place a higher value on perceived usefulness and less value on perceived ease of use, when compared to a general population.

The additional contribution of perceived system performance to the model in predicting perceived usefulness was found to be not significant, although a positive correlation was found between perceived system performance and perceived usefulness. This is perhaps not too surprising, as TAM informs that perceived ease of use is a significant predictor of perceived usefulness. For example, pen and paper has emerged as the more dominant technology in terms of usefulness, when compared to a clay tablet and scribe. It is more likely that the impact of ease of use, rather than system performance, has resulted in the widespread adoption of pen and paper.

The number of dialogue failures was not shown to be a significant predictor of system performance. One might expect a failed dialogue to impact on the user perception of system performance, however, from the 260 tasks executed by clinicians during the study, only six (2.30%) ended in dialogue failure. This relatively small percentage may explain why the perception of system reliability, dependability and accuracy was not significantly affected by a clinician being unable, or unwilling, to complete a task because of dialogue failure.

The perceived speed of dialogue in terms of user turns was found to be a significant predictor in perceived ease of use, but perceived system response time was found not to be significantly correlated with perceived ease of use. TAM defines

Table 3. Additional measures of perceived system quality

<table>
<thead>
<tr>
<th>Question</th>
<th>Median</th>
<th>Quartile Deviation</th>
<th>Z-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 – Became impatient using the system</td>
<td>−1.50</td>
<td>1.13</td>
<td>*−2.40</td>
</tr>
<tr>
<td>22 – Clinical narrative natural</td>
<td>1.50</td>
<td>0.75</td>
<td>*2.40</td>
</tr>
<tr>
<td>23 – Work in a way I like to work</td>
<td>2.00</td>
<td>1</td>
<td>**2.69</td>
</tr>
<tr>
<td>24 – Prefer SDS over paper form</td>
<td>2.50</td>
<td>1</td>
<td>**2.72</td>
</tr>
<tr>
<td>25 – Prefer SDS over keyboard and mouse</td>
<td>3</td>
<td>0.63</td>
<td>**2.75</td>
</tr>
</tbody>
</table>

\( n = 10 \)

*\( p < 0.05 \)

**\( p < 0.01 \)

*aRefer to appendix for a complete description of each question

bThe semi-interquartile range
perceived ease of use as the degree to which a person believes that using a system would be free of effort [2]. Interaction time with the system through dialogue is therefore more likely to impact on this perception, rather than the perceived time it takes for the system to respond.

For example, the initiator of conversation A in figure 3 may view the dialogue as coherent, but the conversation could nevertheless be thought difficult and convoluted. With conversation B, however, the interaction is free of effort, or easy, even though there may be a delay in responding. Extended dialogues are sometimes unavoidable, as in the phrase ‘ulcer found’. Here, an additional qualifier is required in order for the system to position the finding. However, some dialogues may become extended due to recognition errors which then require clarification, like conversation A in figure 3. A robust dialogue manager permits high transaction success to be achieved in spite of recognition or understanding errors [9], but at a cost in terms of perceived ease of use.

The maximum dialogue turn per session was found to be a significant predictor of perceived speed of dialogue in terms of user turns, but no significant correlation was found between the mean number of user turns per session and the perceived speed of dialogue. It appears, therefore, that extreme experiences (unusually long dialogues) have more of an impact than average experiences (the average length of a dialogue). Although this demonstrates extended dialogues can have a significant impact on users’ perception of the time taken to complete a task, clinicians did not become impatient when using the system during the evaluation.

The additional contribution of maximum dialogue turns per session in predicting perceived ease of use was not significant, although these variables were significantly correlated. This indicates that variability in perceived ease of use shared with maximum dialogue turns per session was mediated by clinicians perception of speed of dialogue in terms of user turns.

Recent longitudinal research on technology acceptance [13] has studied the effects of anchors (general beliefs about computers and computer usage) and adjustments (beliefs that are based on direct experience with using the target system) on perceived ease of use. The following anchors were found to affect perceived ease of use: self-efficacy, perceptions of control and computer anxiety. With increasing experience in system use, additional predictors were adjustments in the form of objective usability and perceived enjoyment. Another longitudinal study [10] identified social influences (including subjective norm) and cognitive instrumental processes (including result demonstration, perceived ease of use, interaction job relevance and output quality) as predictors of perceived usefulness over time. Future research into the effects

<table>
<thead>
<tr>
<th>Turn</th>
<th>Conversation A</th>
<th>Turn</th>
<th>Conversation B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is the weather like outside? Did you say the weather?</td>
<td>1</td>
<td>What is the weather like outside?</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td></td>
<td>&lt;delay in responding&gt;</td>
</tr>
<tr>
<td>3</td>
<td>Could you say that again?</td>
<td>4</td>
<td>It is fine and sunny</td>
</tr>
<tr>
<td>5</td>
<td>Yes, the weather outside</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>It is fine and sunny</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Two possible conversations about the weather
of the predictors identified in these studies in relation to the use of a SDS would enhance theoretical understanding of how a set of interrelated drivers and inhibitors affect the acceptance of SDS technology and could further inform its design.

5. Summary

Previous attempts to computerize the recording of observations at the end of an examination have failed, because they were not adopted and used by clinicians. In addition, both paper-based and computer systems using keyboard and mouse-driven metaphors are unable to record information at source. A Spoken Dialogue System (SDS) is by design a more natural and appropriate way of recording clinical observations during an endoscopic examination. This study has demonstrated a high level of user acceptance and clinicians want to use spoken dialogue technology for recording clinical observations during an endoscopic examination. Clinicians would also prefer to use a SDS for recording endoscopy, rather than use a paper-based or keyboard and mouse-driven system. Using a clinical narrative was also perceived to be a natural way to record endoscopic findings. Relationships between basic TAM variables were confirmed and relationships between quality of dialogue measures and TAM variables were established.

References
Appendix – Evaluation Questionnaire Items

All questionnaire items used a seven point Likert scale, with end points strongly agree (3) to strongly disagree (−3) and a neutral score of 0.

(1) The spoken dialogue system enabled me to record findings accurately.
(2) Using the spoken dialogue system is a reliable way to record endoscopic findings.
(3) The spoken dialogue system was dependable.
(4) The spoken dialogue system recorded a finding without me having to repeat the observation.
(5) The spoken dialogue system immediately recorded my observation without asking any further questions.
(6) When making an observation, the spoken dialogue system requested appropriate information before recording the finding.
(7) I found the spoken dialogue system slow to respond after making my observation.
(8) I found the spoken dialogue system quick to respond after making my observation.
(9) I became impatient when using the spoken dialogue system to record findings.
(10) I found it easy to record clinical findings using the spoken dialogue system.
(11) When interacting with the spoken dialogue system I found it clear and understandable.
(12) I found it easy to get the spoken dialogue system to do what I wanted it to do.
(13) Using the spoken dialogue system did not require a lot of my mental effort.
(14) Using the spoken dialogue system would support critical aspects of my job.
(15) I would find the spoken dialogue system useful during an endoscopic examination.
(16) Using the spoken dialogue system would enhance my effectiveness during an endoscopic examination.
(17) Using the spoken dialogue system would increase my productivity during an endoscopic examination.

Introduction to questionnaire items 18 to 25:

‘The current spoken dialogue system only allows you to record ulcer findings for the upper gastrointestinal tract. A spoken dialogue system used in a live endoscopic encounter would enable you to record any type of finding for both the upper and lower gastrointestinal tracts. It would also enable you to record clinical procedures, for example, taking a biopsy. Given this complete system for recording observations and procedures during endoscopy in a live environment, please answer the following questions’

(18) I intend to use the spoken dialogue system to record findings and procedures during endoscopy.
I want to use the spoken dialogue system to record findings and procedures during endoscopy.

I would like to use the spoken dialogue system to record findings and procedures during endoscopy.

It is likely that I would use the spoken dialogue system to record findings and procedures during endoscopy.

I found that using a clinical narrative was a natural way to record clinical findings.

Using the spoken dialogue system to record endoscopic findings enables me to work in a way I like to work.

I would prefer using the spoken dialogue system for recording endoscopic observations rather than using a paper form at the end of an examination.

I would prefer using the spoken dialogue system for recording endoscopic observations rather than using a keyboard and mouse at the end of an examination.