Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)*

KATE S. HONE

Department of Information Systems and Computing,

Brunel University,

Uxbridge,

Middlesex,

UB8 3PH,

United Kingdom.

email: Kate.Hone@brunel.ac.uk

ROBERT GRAHAM

Personal Networks Group,

Motorola.

Midpoint,

Alencon Link,

Basingstoke,

Hants,

RG21 7PL,

United Kingdom.

email: robert_graham@europe27.mot.com

* The first author was supported in this research by EPSRC grant reference GR/L94710. The second author carried out research while at the Husat Research Institute, Loughborough University as part of the SPEECH IDEAS project, funded by the ESRC and DETR under the UK government's LINK IST programme.

Abstract

Applications of speech recognition are now widespread, but user-centred evaluation methods are necessary to ensure their success. Objective evaluation techniques are fairly well established, but previous subjective techniques have been unstructured and unproven. This paper reports on the first stage in the development of a questionnaire measure for the Subjective Assessment of Speech System Interfaces (SASSI). The aim of the research programme is to produce a valid, reliable and sensitive measure of users' subjective experiences with speech recognition systems. Such a technique could make an important contribution to theory and practice in the design and evaluation of speech recognition systems according to best human factors practice. A prototype questionnaire was designed, based on established measures for evaluating the usability of other kinds of user interface, and on a review of the research literature into speech system design. This consisted of 50 statements with which respondents rated their level of agreement. The questionnaire was given to users of four different speech applications, and Exploratory Factor Analysis of 214 completed questionnaires was conducted. This suggested the presence of six main factors in users' perceptions of speech systems: System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability and Speed. The six factors have face validity, and a reasonable level of statistical reliability. The findings form a useful theoretical and practical basis for the subjective evaluation of any speech recognition interface. However, further work is recommended, to establish the validity and sensitivity of the approach, before a final tool can be produced which warrants general use.

1 Introduction

After many years of failing to make its predicted breakthrough, speech recognition technology is now beginning to find its way into peoples' everyday lives. Speech input is in widespread use for applications such as telephone answering services (e.g. Beacham & Barrington, 1996), PC dictation (e.g. Taylor, 1999), over-the-phone travel enquiries (e.g. Failenschmid & Thornton, 1998), and in-car systems (e.g. Howard, 1998). However, the increasing number and variety of inexperienced users of this technology heightens the importance of designing interfaces according to good human factors principles, to ensure their usability. A fundamental precept of the discipline of human factors is to involve the user in all stages of a system's design, from concept to product. Without a user-centred approach, systems can all too easily be developed which are inappropriate or inefficient, or ultimately, which lead to such low levels of user acceptance that they are rejected completely.

This paper is concerned with usability evaluation of interfaces for speech input systems. This refers to systems which allow user input via voice using automatic speech recognition technology and includes a wide range of different types of system, from those which accept only a very limited set of spoken command words to those which accept a sub-set of spoken natural language. These systems also vary in the way that they communicate to the user (e.g. by speech output, visual output or simply carrying out a command). Speech input is taken as the key defining feature here because the probabilistic nature of the recognition process clearly differentiates these systems from most other modes of human-computer interaction.

In general, measures of a system's usability can be defined as objective or subjective. Objective measures, such as task completion time, number of errors, or physiological changes in the user (e.g. heart rate variability) can of course be extremely useful in speech system design and evaluation. A number of useful objective measures are discussed in Gibbon, Moore and Winski (1998). However, these must also be supported by subjective measures to examine user acceptance.

Popular subjective evaluation techniques include open interviews or focus groups. Such qualitative techniques have the advantages of providing a wealth of information, and insights into aspects of system acceptance that could not be predicted prior to data collection. However, subjective measures need not be any less structured or quantifiable than objective measures. Questionnaires, user-completed rating scales, structured interviews, and expert checklists can all produce ‘hard’ data. Any measurement technique, whether objective or subjective, should have the fundamental characteristics of (Sanders & McCormick, 1993):

- reliability (the results should be stable across repeated administrations)
- validity (the technique should measure what it is really intended to measure)
- sensitivity (the technique should be capable of measuring even small variations in what it is intended to measure)
- freedom from contamination (the measure should not be influenced by variables that are extraneous to the construct being measured)

1.1 Generic subjective usability evaluation methods

Before discussing subjective measures that are specific to speech systems, it is helpful to consider more general subjective usability evaluation methods. Two of the most well known questionnaire measures are Shneiderman’s Questionnaire for User Interaction Satisfaction (‘QUIS’ Shneiderman, 1998) and the Software Usability Measurement Inventory (‘SUMI’ Kirakowski, 1996) developed by the Human Factors Research Group at University College Cork. These measures also illustrate two important approaches to questionnaire design.

Shneiderman's QUIS measure was based on his theoretical model of what makes software usable. It consists of one scale to measure overall reactions to the software and then four scales designed to evaluate user reactions to the display, terminology and system information, learning and system capabilities. Sections have recently been added to the questionnaire on multimedia (though not including speech input) and teleconferencing (Shneiderman, 1998).

In contrast to QUIS, SUMI was developed without an a priori expectation of what factors make up usability. Instead the developers produced a large bank of questions, such as ‘this software responds too slowly to inputs’, and gave these to a sample of computer users. Factor analysis was then used to determine the main components of user attitude, and measurement scales for each of these components were further developed in an iterative process. This is an established method for the development of psychometric instruments and has the advantage of reflecting user experience with the software, rather than simply developer expectations. The main sub-scales of SUMI are Affect (or Likeability), Efficiency, Helpfulness, Control and Learnability. Studies have been carried out to support both the validity and reliability of SUMI (Kirakowski, 1996). The developers are currently working towards similarly structured questionnaire tools to assess the usability of web sites (‘WAMMI’) and multimedia software (‘MUMMS’).

Although these and similar techniques have been found to be useful in evaluating a variety of applications (Kirakowski, 1996; Shneiderman, 1998), they are not claimed to be applicable to speech recognition interfaces. Speech systems have a number of unique features that are not addressed within general software usability scales such as SUMI or QUIS. Most importantly, all speech recognisers make errors, and consequently need to give the user feedback as to what has been recognised. The question of how accurate a recogniser must be, while still remaining useful and acceptable, is one that is crucial to industry's development of speech applications. Speech interfaces are also unusual in that users tend to have strong pre-conceived ideas (from human-human conversation) about how an interaction should proceed. Therefore, questions of naturalness, intuitiveness or 'habitability' are important, and are not covered in sufficient depth in general scales.

1.2 Subjective usability evaluation methods specific to speech systems

Because of the lack of validated methods for the subjective evaluation of speech systems, previous research studies have tended to use piecemeal techniques. Two of the less structured methods are the use of open interviews or overall rating scales. For example, Nelson (1986) asked users what they thought of a novel voice recognition system in a product inspection environment, and noted comments such as 'at first it was kind of strange and almost like you were sitting there talking to yourself, but once we got used to it and I started working with it full time, it was a lot faster'. Brems, Rabin & Waggett (1995) studied prompt design for an automated operator service, and experimental participants were asked to rate system options as poor, fair, good or excellent. The authors reported that approximately sixty per cent of users rated a question-plus-options system as excellent, whereas only thirty per cent rated an options-only condition as excellent. This level of data is extremely limited, and does not really allow the designer to improve the system (what exactly is it about the user interface that makes it seem good or bad?).

A more structured method is the use of adjective pairs within rating scales, as tested by Dintruff, Grice & Wang (1985) and Casali, Williges & Dryden (1990). Dintruff *et al*. (1985) examined acceptance of speech systems via 20 adjective pairs, each rated on a ten-point scale with labelled end-points. The twenty pairs consisted of ten ‘feeling items’ and ten ‘attitude items’, as shown in Table 1.

(INSERT TABLE 1 ABOUT HERE)

The same twenty adjective pairs were used to rate the voice input and voice output aspects, separately, of an office speech system supporting functions such as diary keeping and call management. Overall feeling measures were calculated as the mean of the ten individual feeling scores, and similarly overall attitude measures as the mean of the ten attitude scores. The authors used the technique to compare ratings before and after using the system, finding that respondents developed more favourable attitudes to the technology after having used it.

Casali, Williges & Dryden (1990) used thirteen bipolar adjective rating scales of seven intervals each. These consisted of an overall acceptability scale (Acceptable / Unacceptable) and twelve others, as shown in Table 2.

(INSERT TABLE 2 ABOUT HERE)

The ratings were coded into a numerical range between one and seven, and the twelve scores were summed to give a single measure of acceptability, referred to as the Acceptability Index (AI). Casali *et al*. found that the scores on each of the twelve scales were highly correlated with the Acceptable/ Unacceptable scale. The AI score was then used by the authors to show that recognition accuracy was a more important predictor of acceptability than available vocabulary for a data entry speech system. They also noted that older subjects consistently rated speech recognition systems more favourably than younger subjects. The same scale was used by Dillon, Norcio & DeHaemer (1993) who found an effect of subject experience on the AI score, but no effect of vocabulary size.

Zajicek (1990) used a questionnaire format, deriving her questions from CUSI (The Computer User Satisfaction Inventory, an early version of SUMI) and a scale developed by Poulson (1987) to rate the perceived quality of software interfaces. Items were taken from these questionnaires 'where it was felt appropriate', leading to a thirty-item questionnaire, with ten general items concerning the speech interface, and nine concerning the specific prototype interface. Each question was worded as a statement, to which users responded on a scale from -3 (disagree strongly) to +3 (agree strongly). Examples of the general statements included 'The equipment is confusing to use', 'I have to concentrate hard to use the equipment' and 'A speech interface is easier than a keyboard'. The results were not subjected to any statistical analysis; rather, Zajicek based her conclusions on a comparison of the absolute scores between 3 user groups. Interestingly, Zajicek also carried out interviews with subjects investigating what factors they considered to be important in a speech system. She concluded that four evaluation areas - controllability, user satisfaction, learnability and technical performance (in order of priority) - should be used to provide a framework for future evaluations.

Kamm, Litman & Walker (1998) tested a user satisfaction survey, with the ten questions shown in Table 3.

(INSERT TABLE 3 ABOUT HERE)

There were five possible responses to most of the questions (labelled 'almost never' / 'rarely' / 'sometimes' / 'often' / 'almost always' or an equivalent range), but some questions just had three responses ('yes' / 'no' / 'maybe'). The responses were mapped to integer values between one and five. A Cumulative Satisfaction score was calculated by summing the scores for each question. Kamm *et al*. found that three variables, perceived task completion, mean recognition score, and number of help requests, were significant predictors of this cumulative satisfaction score.

Finally, researchers at the University of Edinburgh, in collaboration with British Telecom, have also used questionnaires to evaluate over-the-phone services incorporating speech input. A number of versions of the questionnaire have been reported, all originally based on Poulson's (1987) indices. For example, Love (1997) lists thirty-two attitude statements such as 'I found the (system) easy to use', 'I had to concentrate hard when using the (system)', and 'I thought the (system) was reliable'. Many of the same questions appear in a shorter twenty-two item questionnaire reported in Dutton, Foster & Jack (1999). Each of these questionnaires use five- or seven-point scales (labelled 'strongly agree' to 'strongly disagree' with a point marked 'neutral'). The general approach of this research group is to calculate an average score based on the responses of all the attitude questions. When comparing more than one system they also look for differences in the mean ratings given on individual questions (e.g. McInnes, Nairn, Attwater, & Jack, 1999). Support for the validity of the overall measure is not strong, with Foster *et al*. (1998) reporting important discrepancies between user attitude to three versions of a system as measured by a total score on the questionnaire and an objective measure of user preference. The sensitivity of the measure also seems to be low, with McInnes *et al*. (1999) failing to find any difference in overall ratings given to perfect, intermediate and low accuracy versions of the same interface (and this result was obtained, despite using a large sample size in the experiment).

With some minor exceptions, all of the previously used techniques for subjective speech interface evaluation, outlined above, suffer from the same weaknesses.

First, their content and structure are, for the most part, arbitrary. The items chosen for a questionnaire or rating scales are based neither on theory nor on well conducted empirical research; rather, they are picked by the researchers according to ‘what seems right at the time’. Similarly, the reasons for choosing a particular structure (e.g. questions, statements or numerical scales) and sub-structure (presentation, number of points on a scale, etc.) are not reported.

Second, the techniques have not been satisfactorily validated, either against other subjective measures or against objective measures. There is often no reason to assume that user responses are really measuring the construct of acceptability, rather than some other factor. Also, it is unlikely, given the arbitrary way items have been chosen, that they sample all the facets of acceptability rather than just a limited subset.

Third, there are no reports of the reliability of the techniques used. There are two main types of reliability. The first is test-retest reliability, referring to the stability of the measure over time and found by calculating the correlation between sets of scores measured for the same system on two occasions. The second type of reliability refers to the internal consistency of a measure and is calculated using Cronbach's alpha. The test is applied to the scales that together are thought to measure a particular theoretical construct, such as user satisfaction.

Fourth, the way that the collected data is used is inappropriate. In many of the above examples, scores on individual questionnaire items are simply summed or averaged to give an overall acceptability score. Such an approach can only be justified on the basis of evidence that all of the items are measuring the same construct, otherwise the overall score will be meaningless. The individual items may represent different constructs; one cannot simply add chalk to cheese. The alternative approach of comparing systems on the basis of scores on individual questionnaire items is also problematic because people are likely to vary in the way that they interpret the item wording. Well designed measures of attitude should always include a number of items, all mapping onto the same construct, in order to overcome variability in the measure due to extraneous features of this kind.

It can be concluded that none of the existing techniques for subjective speech interface evaluation meet the criteria for a valid psychometric instrument. Claims made on the basis of these existing measures (for instance, that a design parameter does or does not affect user attitude) should therefore be treated with a great deal of caution.

1.3 The SASSI approach

Given the shortcomings of existing measures for the subjective evaluation of speech interfaces there is clearly a need for the development of a more valid and reliable approach. Such a technique would have significant benefits for both theory and practice in the development of speech systems. A major benefit is that it would allow meaningful comparisons be made between alternative interfaces. In addition it could be used in benchmarking for new product development.

The current paper describes the first step towards the development of such a tool for the Subjective Assessment of Speech System Interfaces (SASSI). The ultimate aim of the research is to produce a subjective tool that is:

- valid, reliable and sensitive
- widely applicable to all types of speech recognition interface, from command and control to data entry and interactive database enquiry applications
- quickly and easily completed by naïve and/or first-time respondents
- quantifiable, to allow statistical comparison of multiple interface options or benchmarking of a single option
- discriminative, to allow identification of the good and bad aspects of a design, and inherently suggest possible remedies
- complete, capturing all the important aspects of a user's experience with a speech recognition system.

In order to meet the requirement for a quick and quantifiable method it was decided that the measure should be in the form of a questionnaire to be completed by users of the system. In the absence of firm theoretical guidance on the features of speech systems that contribute to user satisfaction, it was decided to use an empirical approach to develop the questionnaire. As in the development of SUMI discussed above (Kirakowski, 1996) this involves generating a large pool of initial questionnaire items and using empirical methods to determine latent structure from the pattern of users' responses to these questions. The development of a useful measurement tool using this approach involves a process of iterative refinement. This paper describes the first in a series of planned iterations of the design of SASSI. During this stage the emphasis is upon establishing the main components that make up a user's perception of a speech input system and producing reliable scales to measure each of these. This work plays a vital role in laying a solid foundation for future research to address important theoretical questions such as which system characteristics affect user responses and which user responses predict eventual system acceptance. These future research issues are discussed further at the end of the paper.

2 Method

2.1 Questionnaire Item Generation

A decision was taken to use Likert scales; declarative statements of opinion (e.g. 'this system is easy to use') with which respondents rate their agreement, typically with five- or seven-point scales. This method was chosen over the alternative of bipolar adjectives, for two main reasons. The first is that it can sometimes be difficult to determine appropriate opposites for each end of a bipolar scale (for example, in Casali *et al*.'s questionnaire - see Table 2 - is 'facilitating' really the opposite of 'distracting'?). The other reason is that a finer grain of meaning is possible in the items. For instance it is not clear that potentially useful questions, such as 'a high level of concentration is required when using this system', could be converted into simple adjectival descriptors without losing much of their meaning.

An initial pool of attitude statements was generated, based on the general usability questionnaires reviewed in section 1.1 above, and the specific speech measures outlined in section 1.2. A general review of the speech system usability literature (e.g. Baber & Noyes, 1993) suggested a number of additional items which were not specifically addressed in the previous methods. Finally, extra items were added according to the authors' practical experiences of designing and evaluating speech system interfaces. Using this approach we hoped to sample all relevant facets of user opinion and thus ensure the content validity of the measure.

Care was taken to balance the number of positive and negative statements, and duplicated items were removed from the overall pool of statements. A third expert in speech interface usability checked the statements for clarity of meaning, and obviously confusing items were removed. Some potentially problematic items (e.g. 'the interaction with the system is distracting') were retained because they had formed part of previous speech usability questionnaires.

This process of item generation produced a pool of 52 statements. These were ordered in the questionnaire so that positive and negative items were randomly distributed (to prevent respondents being tempted to simply mark straight down a column of responses). Seven-point Likert scales were used, labelled strongly agree, agree, slightly agree, neutral, slightly disagree, disagree, and strongly disagree.

An initial pilot test of the questionnaire revealed that two questionnaire items (both referring to obtaining ‘services’ from the system) could not be generalised to all speech input systems (e.g. command-and-control applications). These items were dropped, resulting in a fifty-item questionnaire.

2.2 Sample and Procedure

226 completed questionnaires were returned over the course of four separate studies involving a total of eight different speech input systems. The choice of applications used was largely pragmatic but was intended to capture a range of different speech input system types. Systems can be categorised in several different ways, for example by the degree of lexical, syntactic and semantic constraint on user utterances, by degree of user / system initiative, by the mode of system output. While we were not able to capture all combinations of these system variables, some of the main contrasts were included as illustrated by the descriptions below.

Study one (Graham, Carter, & Mellor, 1998) involved a small vocabulary system (<20 words) with a strict syntax where the interaction was initiated by the user. Here two versions of a speech interface for dialling telephone numbers were tested. In both versions dialling was accomplished by operating a press-to-talk button, then speaking a command word (‘phone’), the digits in chunks of any size (e.g. ‘01509’-’611’-’0’-’88’), and another command word (‘dial’). One version used audio-plus-visual feedback of the recognition results, and the other audio-only feedback. Forty-eight completed questionnaires were collected for each version.

Study two involved a mixed initiative, medium sized vocabulary system (~100 words) with a syntax which allowed some variation in command structure. Twenty-two participants used an in-car speech interface to operate a variety of features including the car-phone, entertainment system, and climate control. The interactions were a mixture of basic commands (e.g. ‘climate control temperature twenty degrees’) and two-way dialogues (e.g. ‘phone store 01509-611088’ - <‘name please’> - ‘Bob’). Each participant completed the SASSI questionnaire, having experienced the system for the first time over a two-hour session.

Study three used a study with similar parameters to that in study two. Two versions of a voice operated stereo system (encompassing radio, tape and CD functions) were tested. Valid commands included ‘tape reverse’, ‘CD play disc 3 track 5’ and ‘radio tune 97.9 FM’. In one version, explicit audio-plus-visual feedback of the recognition results was given, and in the second, only implicit or ‘primary’ feedback (i.e. the operation of the tape, radio or CD itself) was present. Thirty-two completed questionnaires were collected for the implicit-feedback interface and thirty-one for the explicit-feedback version.

Study four (Hone & Golightly, 1998) involved three versions of an over-the-phone banking application for checking balance, transferring funds, etc. All three were interactive dialogues (with speech input and output) initially initiated by the system. They were explicitly designed to differ in the degree of constraint the system prompts implied over user utterances. At one extreme was a ‘yes/no’ dialogue style where users were asked a series of questions such as “do you want to hear your balance” and were expected to respond with a yes or no answer. At the other extreme was an open query style of dialogue where users were asked open-ended questions such as “which service do you require” and were expected to reply with a limited subset of natural language. Between these extremes was a menu style dialogue where users were given a choice of responses to choose after each prompt (e.g. “which service do you require, balance, cash transfer or other?”). Fifteen completed questionnaires were collected for each version.

All participants in the trials were recruited from the general UK population through advertising. None were experienced users of speech input systems and a range of experience with computers was represented (from complete novice to expert). They were paid between UK£15-30 for participating in the studies.

3 Analysis and Results

3.1 Data Screening

Prior to analysis the data was examined for accuracy of data entry, missing values, and fit between the distributions of the variables and the assumptions of multivariate analysis. This stage in the analysis is very important as problems here can have a large impact on the factor solution obtained.

Six variables were identified which had missing data for greater than five per cent of the ‘cases’ (or respondents). The six questions all referred to the system ‘messages’ or system ‘voice’, and the missing data was due to the inclusion of a sample of respondents who had used a speech system without explicit feedback. It was decided to remove these variables from the analysis in order for the questionnaire to be applicable to all speech-input systems. One case was identified with missing data on twenty-seven per cent of the items, and was removed. A further forty-one missing data points were identified. As these were scattered through the data set, with no apparent pattern, it was decided to replace these with mean values (calculated from the remaining cases with that specific system).

Univariate and multivariate outliers in the data were identified and dealt with as recommended by Tabachnick and Fidell (1996). Eleven cases (respondents) were removed from the analysis at this stage. Skew, Kurtosis and linearity was also assessed and found to be satisfactory.

Following initial data screening, 214 cases remained in the sample and forty-four variables were retained for analysis.

The correlation matrix was examined to check that the requirements for factor analysis were met. Several correlations of .30 or over were observed suggesting the data was suitable. Furthermore the Kaiser-Meyer-Olkin (KMO) test of sampling adequacy gave a result of .95, indicating that the associations between the variables in the correlation matrix can be accounted for by a smaller set of factors (Dziuban & Shirkey, 1974). Bartletts test of Sphericity (BS) was also significant at $p < 0.001$, indicating that there are discoverable relationships in the data.

3.2 Exploratory Factor Analysis

An initial principal components extraction with Varimax rotation was performed using the 'SPSS FACTOR' software tool on the forty-four questionnaire item scores for the sample of 214 cases. The analysis yielded eight factors with Eigenvalues greater than one. Examination of the factor pattern matrix revealed several variables which did not load on any factor (with a criterion of 0.4 to accept a variable as defining a factor , Ferguson & Cox, 1993). In addition, a number of variables were cross-loaded (loading at 0.4 on two or more factors). Following the advice of Ferguson and Cox (1993), non-loading and cross-loading variables (where the difference in magnitude between loadings is less than 0.2) were removed. These items were: (non-loading items) 'the interaction with the system is logical', 'the interaction with the system is natural', 'the interaction with the system is distracting'; and (cross-loading items),'too many steps are required to complete a task with the system', 'the interaction with the system is complicated', 'I sometimes felt angry using the system', 'I felt inhibited speaking to the system', 'I was able to be spontaneous using the system', 'I would prefer to speak to a human operator'. Another iteration of this process led to the removal of a further variable from the analysis: 'I felt comfortable using the system'. Inspection of these removed items shows many of them to be potentially ambiguous, or likely to be affected by social desirability, providing further justification for their removal (Ferguson & Cox, 1993).

Principal components extraction with Varimax rotation on the remaining thirty-four variables produced six factors with Eigenvalues greater than one. The six factor solution was further supported by examination of the Scree Plot and of the residual correlation matrix for three-, four- and five-factor solutions. Communality values were all acceptable (greater than or equal to 0.4) indicating that the variables were well defined by the six factors extracted. While statistical properties, such as these, must be considered when evaluating a factor solution, it is also important to consider the criterion of interpretability. This relies upon the judgement of the analyst. In this case the three-, four- and five-factor solutions were inspected by both authors and none were found to be as readily interpretable as the six factor version.

Table 4 presents the results of the factor analysis. The six factors are listed in order of importance (determined by Eigenvalue magnitude and proportion of variance explained). Only factor loadings greater than 0.45 are shown, in order to increase clarity. Overall, the factor solution accounts for 64.7 per cent of the total variance.

(INSERT TABLE 4 ABOUT HERE)

3.3 Factor Naming

A factor name should capture the underlying dimension which unifies the group of variables loading on that factor (Tabachnick & Fidell, 1996). Both authors independently inspected the items loading on to each factor, with the aim of reducing some of the subjectivity associated with factor naming. The most strongly loading items were deemed most important when interpreting each factor. Where initial namings did not agree, a process of brainstorming was carried out until agreement was reached.

Factor 1 contains items such as ‘the system is accurate’ and ‘the system didn't always do what I wanted’. The items all clearly relate to whether the system recognises users input correctly and hence does what the user intended and expects. We have named this factor ‘System Response Accuracy’.

Factor 2 contains items such as ‘I enjoyed using the system’, ‘The system is friendly’ and ‘I would use this system’. These items are reminiscent of the SUMI dimension of Affect / Likeability. We have chosen the term ‘Likeability’ because the factor includes statements of opinion about the system as well as feeling (affect) items.

Factor 3 contains items such as ‘I felt tense using the system’ and ‘A high level of concentration is required when using the system’. The items seem to summarise both the perceived level of effort needed to use the system and user feelings arising from this effort. We considered a number of names for this factor including stress and mental workload, but agreed on the term ‘Cognitive Demand’.

Factor 4 contains items such as ‘The interaction with the system is repetitive/boring/irritating’. We have named it ‘Annoyance’.

Factor 5 contains items relating to whether the user knows what to say and knows what the system is doing. This could be seen to relate to the concept of ‘visibility’; that is, whether the conceptual model of the system, the alternative actions and the results of these actions are visible in the interface (Norman, 1988). However, as the term visibility is clearly unsuitable for those systems without visible output we have chosen the term ‘Habitability’ instead. A habitable system may be defined as one in which there is a good match between the user's conceptual model of the system and the actual system.

Factor 6 contains only two items, both relating to the speed of the system. We have therefore named this factor 'Speed'. Note that one should normally be suspicious of any factor defined by only two items. However, the high loadings (> 0.7) of both variables onto this factor suggest that this factor is viable.

3.4 Sub-scale Reliabilities

The internal consistency reliability of the items loading on each of the six factors defined above was tested using Cronbach's alpha. The internal consistency estimates of the factors were: (1) System Response Accuracy, alpha = 0.90; (2) Likeability, alpha = 0.91; (3) Cognitive Demand, alpha = 0.88; (4) Annoyance, alpha = 0.77; (5) Habitability, alpha = 0.75; (6) Speed, alpha = 0.69. Igbaria and Parasuraman (1991) suggest that alpha values greater than 0.70 are adequate in the early stages of research on hypothesised measures of a construct; all the sub-scales (except Speed, defined by only two variables) meet this criterion. Reliabilities of 0.80 or more are generally required for widely used scales (Igbaria & Parasuraman, 1991) and the System Response Accuracy, Likeability and Cognitive Demand sub-scales all meet this criterion.

4 Discussion

The current paper has reported on the first in a number of planned iterations in the development of SASSI. Exploratory factor analysis on the initial bank of questionnaire items has suggested six main factors which contribute to the user's experience of speech input systems. We have tentatively named these System Response Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability and Speed. 'System Response Accuracy' refers to the user's perceptions of the system as accurate and therefore doing what they expect. This will relate to the system's ability to correctly recognise the speech input, correctly interpret the meaning of the utterance and then act appropriately. This factor accounts for the greatest proportion of the variance in the solution obtained, suggesting that it is a particularly important or salient aspect of a user's interaction with a speech recognition system. The importance of this factor confirms our expectation that generic subjective measures (such as SUMI or QUIS) are unsuitable for the evaluation of speech recognition systems.
'Likeability' refers to the user's ratings of the system as useful, pleasant and friendly. It is similar to the SUMI construct of Affect / Likeability, suggesting that this factor generalises across speech and non-speech input software. 'Cognitive Demand' refers to the perceived amount of effort needed to interact with the system and the feelings resulting from this effort. 'Annoyance' refers to the extent to which users rate the system as repetitive, boring, irritating and frustrating. The emergence of this as a separate factor from Likeability is interesting and, if confirmed in future work, may also suggest a difference between speech and non-speech input systems. 'Habitability' refers to the extent to which the user knows what to do and knows what the system is doing. It can be understood in terms of the adequacy of the user's conceptual model of the speech system as a dialogue partner (Baber, 1993). It is likely that the more complex the system, the more important this factor may become as users struggle to

understand the limits of the system (lexical, syntactic, semantic and pragmatic). Finally, ‘Speed’ refers to how quickly the system responds to user inputs.

The emergence of an underlying structure in the questionnaire response set confirms our expectation that user attitude to speech recognition systems is not a unidimensional construct. This finding further calls into question the usefulness of measures produced by summing or averaging user responses to ad hoc questionnaires. However, further research is needed to confirm the specific factor structure presented in this paper. In particular it must be recognised that the questionnaire to date has only been used with a limited range of speech recognition systems. Future research will need to be extended to include many more applications, particularly more examples of complex spoken language dialogue systems.

The current paper has also explored the reliability of the sub-scales loading onto each of the six factors identified. Three of these, System Response Accuracy, Likeability and Cognitive Demand, have reliabilities of more than 0.80, the level required for a scale to be considered acceptable (Igbaria & Parasuraman, 1991). However, these levels of reliability need to be confirmed with a statistically independent sample. Two of the scales, Annoyance and Habitability, have reliabilities of more than 0.70, which is considered adequate in the early stages of research (Igbaria & Parasuraman, 1991). Both of these have relatively few items loading onto the factor concerned. Therefore, in future iterations of the questionnaire, it is intended that extra items will be designed with the aim of contributing further to the reliable measurement of those factors. The same is true for the Speed sub-scale, which currently includes only two items and has the lowest reliability of all the sub-scales (alpha value of 0.69).

To date the development of SASSI has concentrated on establishing it as a reliable measure. This work is vital in providing a solid underpinning for future theoretical research. This future work will assess the validity of the measure and, related to this, what the measure really means for designers. From a methodological point of view it is important that the validity of a measure is established. Face validity refers to whether the measure 'looks like' it is measuring what it should. This criterion can be important in getting a measure accepted by other researchers in the field, but generally isn't considered important by measurement experts (Lehman, 1991). We would argue that SASSI has an acceptable level of face validity. It appears to be measuring aspects of interacting with speech input systems which we and others have hypothesised as being important. Construct and predictive validity are more important features of a measurement tool. Construct validity can be established by investigating the degree to which a measure correlates with other measures thought to be measuring similar constructs. In the case of SASSI, construct validity will be investigated through correlation of the sub-scales with established usability scales such as SUMI and QUIS. Predictive validity is central to the eventual success of SASSI. This refers to the degree to which the measure is predictive of external criteria. In this case these criteria might be whether users accept a system or choose the speech system over alternatives. It can be hypothesised that the different SASSI sub-scales will vary in the degree to which they correlate with user preferences or behaviour. If this is the case, regression techniques can be used to determine an aggregate score, based on the individual SASSI measures, giving appropriate weight to each sub-scale. If a score derived from the SASSI sub-scales in this way can be shown to be a significant predictor of behavioural metrics, then it will have important implications. First, it can be used to operationalise the dependent variable in experimental investigations of which features of speech input systems affect user satisfaction.

Second, it can be used to evaluate prototype systems during the development process, hopefully resulting in improvements in design.

It was stated in the introduction that SASSI should be both widely applicable (relevant to all speech recognition applications) and complete (capturing all relevant aspects of a user's subjective experience with the system). During the course of the current research, a conflict arose between these two aims. In attempting to produce a complete measure, a number of items were included which referred to the system 'messages' or system 'voice'. During the use of the questionnaire it became clear that these items were not applicable to the users of one of the speech systems tested, which did not provide any explicit feedback of the recognition results. In order to preserve the broad applicability of the questionnaire, these items were therefore removed from the analysis. However, it might be argued that their exclusion calls into question SASSI's claim for completeness (or 'content validity'). It is therefore proposed that further development of the questionnaire addresses the question of system feedback in more detail. There are two possible approaches that can be followed. The first is to generate questionnaire items that can be meaningfully interpreted by users of systems with either implicit or explicit feedback (and regardless of output modality). This method is preferable from the point of view of producing a generally applicable measure, but may prove impractical due to the difficulty of phrasing appropriate attitude statements. The second alternative is to develop questions that apply to any system using explicit feedback. These could form a separate section of the questionnaire that is only completed if explicit feedback is used. Of course the reliability and validity of either approach would need to be carefully assessed.

A further aim in the future development of SASSI is to produce a more useful and more user-friendly version of the questionnaire. While the improvements to SASSI proposed above would result in a measure which could be used for comparing systems, it would be helpful to provide system developers with a measure to evaluate systems in isolation. In order to do this, a large bank of normative data must be collected from a wide variety of applications and with a wide variety of users (in terms of age, gender, experience, etc.). This can then lead to the development of a scoring system, to judge the relative quality of an interface against the norm. Population norms can also improve the interpretation of results from the tool. For example, if it was found that older people tended to respond more positively using SASSI than younger people (cf. Casali *et al*., 1990), then this should be taken into account each time a group of older users are tested. Improved user friendliness can be accomplished by reducing the length of the questionnaire (i.e. reducing the number of items), and providing background instructions which are helpful and easily understood.

References

Baber, C. 1993. Developing interactive speech technology. In C. Baber & J. Noyes (Eds.), *Interactive Speech Technology* , pp. 1-18. London: Taylor & Francis.

Baber, C., & Noyes, J. M. (Eds.). 1993. *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input / Output to Computers*. London: Taylor & Francis.

Beacham, K., & Barrington, S. 1996. CallMinder™ - The development of BT's new telephone answering service. *BT Technology Journal*, *4*(2), 52-59.

Brems, D. J., Rabin, M. D., & Waggett, J. L. 1995. Using natural language conventions in the user interface design of automatic speech recognition systems. *Human Factors, 37*(2), 265-282.

Casali, S. P., Williges, B. H., & Dryden, R. D. 1990. Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. *Human Factors, 32*(2), 183-196.

Dillon, T. W., Norcio, A. F., & DeHaemer, M. J. 1993. Spoken language interaction: effects of vocabulary size and experience on user efficiency and acceptability. In G. Salvendy & M. J. Smith (Eds.), *Human-Computer Interaction: Software and Hardware Interfaces, Proceedings of the 5th International Conference on Human-Computer Interaction (HCI International '93)* , pp. 140-145. Amsterdam: Elsevier.

Dintruff, D. L., Grice, D. G., & Wang, T. G. 1985. User acceptance of speech technologies. *Speech Technology, 2*(4), 16-21.

Dutton, R. T., Foster, J. C., & Jack, M. A. 1999. Please mind the doors - do interface metaphors improve the usability of voice response services. *BT Technology Journal, 17*(1), 172-177.

Dziuban, C., & Shirkey, E. 1974. When is a correlation matrix appropriate for factor analysis? *Psychological Bulletin, 81*, 358-361.

Failenschmid, K., & Thornton, J. H. S. 1998. End-user driven dialogue system design: the REWARD experience, *Proceedings of the 5th International Conference on Spoken Language Processing* , Vol. 2, pp. 37-40. Rundle Mall, Australia: Causal Productions.

Ferguson, E., & Cox, T. 1993. Exploratory factor analysis: a users' guide. *International Journal of Selection and Assessment, 1*(2), 84-94.

Foster, J. C., McInnes, F. R., Jack, M. A., Love, S., Dutton, R. T., Nairn, I. A., & White, L. S. 1998. An experimental evaluation of preferences for data entry method in automated telephone services. *Behaviour and Information Technology, 17*(2), 82-92.

Gibbon, D., Moore, R. and Winski, R. 1998. *Handbook of Standards and Resources for Spoken Language Systems. Volume 3: Spoken Language System Assessment.* Berlin: Mouton de Gruyter.

Graham, R., Carter, C., & Mellor, B. 1998. The use of automatic speech recognition to reduce the interference between concurrent tasks of driving and phoning, *Proceedings of the 5th International Conference on Spoken Language Processing* , Vol. 4, pp. 1623-1626. Rundle Mall, Australia: Causal Productions.

Hone, K. S., & Golightly, D. 1998. Interfaces for speech recognition systems: the impact of vocabulary constraints and syntax on performance, *Proceedings of the 5th International Conference on Spoken Language Processing* , Vol. 4, pp. 1199-1202. Rundle Mall: Causal Productions.

Howard, K. 1998. Talking to your car. *AutoCar*, 29 July 1998, 40-41.

Igbaria, M., & Parasuraman, S. 1991. Attitudes towards microcomputers: development and construct validation of a measure. *International Journal of Man-Machine Studies*, *35*, 553-573.

Kamm, C. A., Litman, D. J., & Walker, M. A. 1998. From novice to expert: the effect of tutorials on user expertise with spoken dialogue systems, *Proceedings of the 5th International Conference on Spoken Language Processing* , Vol. 4, pp. 1211-1214. Rundle Mall, Australia: Causal Productions.

Kirakowski, J. 1996. The software usability measurement inventory: background and usage. In P. Jordan (Ed.), *Usability Evaluation in Industry* , pp. 169-177. London: Taylor & Francis.

Lehman, R. S. 1991. *Statistics and Research in the Behavioural Sciences*. Belmont, CA: Wadsworth Publishing Company.

Love, S. 1997. *The Role of Individual Differences in Dialogue Engineering for Automated Telephone Services*. Unpublished PhD thesis, University of Edinburgh, Edinburgh, UK.

McInnes, F. R., Nairn, I. A., Attwater, D. J., & Jack, M. A. 1999. Effects of prompt style on user responses to an automated banking service using word-spotting. *BT Technology Journal, 17*(1), 160-171.

Nelson, D. L. 1986. User acceptance of voice recognition in a product inspection environment, *The Official Proceedings of Speech Tech '86: Voice Input/ Output Applications Show and Conference* , pp. 62. New York: Media Dimensions Inc.

Norman, D. A. 1988. *The psychology of everyday things*. New York: Basic Books.

Poulson, D. F. 1987. Towards simple indices of the perceived quality of software interfaces, *Proceedings of the IEE Colloquium on Evaluation Techniques for Interactive Systems Design*. London: Institute of Electrical Engineers.

Sanders, M. S., & McCormick, E. J. 1993. *Human Factors in Engineering and Design*. (7th ed.). New York: McGraw-Hill.

Shneiderman, B. 1998. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* (3rd ed.). Reading, MA: Addison Wesley.

Tabachnick, B. G., & Fidell, L. S. 1996. *Using Multivariate Statistics.* (3rd ed.). New York: Harper Collins.

Taylor, P. 1999. The power of speech in the digital age. *The Financial Times Review of Information Technology (FT-IT Review), 3,* 3.

Zajicek, M. P. 1990. *Evaluation of a speech driven interface.* Proceedings of the UK IT 1990 Conference. Southampton, 19-22 March 1990..

Table 1: Adjective pairs used by Dintruff *et al*. (1985)

FEELING ITEMS	**ATTITUDE ITEMS**
Uncomfortable / Comfortable	Unfavourable / Favourable
Passive / Active	Hard to use / Easy to use
Tense / Relaxed	Unreliable / Reliable
Angry / Friendly	Slow / Fast
Sad / Happy	Useless / Useful
Depersonalized / Individualized	Rigid / Flexible
Bored / Interested	Inefficient / Efficient
Weak / Strong	Worthless / Valuable
Inhibited / Spontaneous	Inaccurate / Accurate
Dissatisfied / Satisfied	Inappropriate / Appropriate

Table 2: Adjective pairs used by Casali *et al*. (1990)

Fast / Slow
Accurate / Inaccurate
Consistent / Inconsistent
Pleasing / Irritating
Dependable / Undependable
Natural / Unnatural
Complete / Incomplete
Comfortable / Uncomfortable
Friendly / Unfriendly
Facilitating / Distracting
Simple / Complicated
Useful / Useless

Table 3: Questions used by Kamm *et al*. (1998) (slightly paraphrased in order to be generalisable)

Did you complete the task?
Was the system easy to understand?
Did the system understand what you said?
Was it easy to find the message you wanted?
Was the pace of interaction with the system appropriate?
Did you know what you could say at each point of the dialogue?
How often was the system sluggish and slow to reply to you?
Did the system work the way you expected it to?
How did the system's voice interface compare to a manual interface?
Do you think you would use the system regularly?

Table 4: Exploratory Factor Analysis Results

	Component					
	1	**2**	**3**	**4**	**5**	**6**
The system is accurate	.799					
The system is unreliable	-.736					
The interaction with the system is unpredictable	-.719					
The system didn't always do what I wanted	-.718					
The system didn't always do what I expected	-.713					
The system is dependable	.696					
The system makes few errors	.674					
The interaction with the system is consistent	.586					
The interaction with the system is efficient	.580					
The system is useful		.698				
The system is pleasant		.668				
The system is friendly		.621				
I was able to recover easily from errors		.606				
I enjoyed using the system		.587				
It is clear how to speak to the system		.578				
It is easy to learn to use the system		.569				
I would use this system		.538				
I felt in control of the interaction with the system		.482				
I felt confident using the system			.746			
I felt tense using the system			-.725			
I felt calm using the system			.699			
A high level of concentration is required when using the system			-.610			
The system is easy to use			.604			
The interaction with the system is repetitive				.757		
The interaction with the system is boring				.684		
The interaction with the system is irritating				.586		
The interaction with the system is frustrating				.509		
The system is too inflexible				(.429)		
I sometimes wondered if I was using the right word					.676	
I always knew what to say to the system					-.609	
I was not always sure what the system was doing					.597	
It is easy to lose track of where you are in an interaction with the system					.597	
The interaction with the system is fast						-.778
The system responds too slowly						.723
Percentage of Variance (rotated solution)	**16.46**	**13.95**	**11.62**	**8.78**	**7.53**	**6.34**