Subjective Assessment of Speech-System Interface Usability

Kate S. Hone and Robert Graham*

Department of Information Systems and Computing, Brunel University, UK; 
*Motorola Internet Software & Content Group, Midpoint, Alencon Link, Basingstoke, Hants, UK.

Kate.hone@brunel.ac.uk; rob.graham@motorola.com

Abstract

Methods for the evaluation of the efficiency and effectiveness of speech input / output systems are well established. However, user subjective reactions to speech interfaces may well be a more important predictor of real world success. A review of existing subjective measures of speech system usability reveals a number of limitations in the approaches taken. This paper then summarises the work we have conducted in developing a new measure for the subjective assessment of speech system interfaces (SASSI).

1. Introduction

While interactive systems using voice input/output have been available for many years, there are several indications that we are now entering a crucial period in their deployment. The US has recently seen a proliferation of 'voice portals' such as Tellme and BeVocal, which offer public access to a wide range of internet content (e.g. news, travel information, weather, shares) via speech recognition and speech output over the telephone network. Several companies in Europe are planning to launch similar products in the near future. Progress is also being made with embedding speech technology into everyday products and even into the fabric of the home. These increasingly consumer oriented applications of speech technology mean that there is a need for valid consumer oriented evaluation methods which consider users' subjective responses to speech systems.

In this paper we review the methods which have been used to assess user satisfaction with speech systems and summarise the work which we have conducted in developing a new measure for the subjective assessment of speech system interfaces (SASSI).

2. Measuring Usability

The ability to accurately measure speech system usability is a vital component for both good system design and theoretical research in this area. Without good measurement we cannot assess whether systems have been improved through redesign, nor can we evaluate whether particular design features affect usability, nor can we compare alternative systems. The concept of usability is generally defined in terms of effectiveness, efficiency and satisfaction [1]. Where the users are consumers (of a product or service) subjective satisfaction is arguably the most important component. Ultimately people will not choose to use systems which they rate poorly. Using subjective measures in the evaluation of such products, especially during development and research, provides a tangible means of putting people at the centre of information technology. In speech system evaluation, measurement of subjective satisfaction is an area which has received little attention.

2.1. The Importance of Subjective Satisfaction with Speech Systems

There are two reasons for a need for particular emphasis on subjective satisfaction with speech systems (compared to more traditional manual input / visual output systems). The first (1) concerns the changing context of use of the technology, the second (2) concerns the nature of the technology itself.

1. Context of use: In the past speech input / output technology was successful only in a limited number of specialised domains. Now speech technology is increasingly seen as a means of widening access to information services. Many see speech as an ideal gateway to the mobile internet. Others see it as a way of encouraging more widespread use of information technology, particularly by previously excluded groups, such as older people. The eventual success of speech as a means of broadening access in this way is very heavily dependent on the perceived ease of use of the resulting systems.

2. The nature of speech technology: Despite great improvements in speech technology over the years, problems do remain. The problems with recognition error in particular mean that we have a technology which, however well designed, has the potential to perform badly. It is therefore vital that we know what level of performance users will tolerate. Objective measures of system performance, though important, are not sufficient.

2.2. Measuring User Satisfaction

When we measure something there are certain qualities we require from our measuring instrument, for instance we expect the instrument to give the same results (when measuring the same thing) on different occasions. Some fundamental characteristics of good measurement are [11]:

- Reliability (the results should be stable across repeated administrations).
- Validity (the technique should measure what it is 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).

These characteristics must be borne in mind when designing or selecting a measuring tool. In any research which involves measurement, the conclusions will always be limited by the quality of the measuring instrument used.

When the quality to be measured is subjective (involving people’s thoughts and feelings) rather than objective, the requirement for scientific rigour in the measuring tool is just as strong, but becomes more difficult to achieve. For example when people are asked to rate their agreement with a statement, subtle variations in wording can have strong effects on ratings, different people may interpret the same statement differently, and ratings can be influenced by a desire to appear "normal" (known as the social desirability effect). The discipline of psychometrics provides methods for developing valid and reliable measurement instruments given these constraints. Typically these measures take the form of a set of questions, attitude statements or adjectives with associated rating scales. Such measures are time consuming and expensive to develop, since large samples of data are needed in order to establish the psychometric properties of an instrument. However, the investment is justified by the improved quality of the resulting tool and the increased confidence you can then place on any results obtained.

Methods drawn from the field of psychometrics have been successfully used in the development of user satisfaction measures, most notably in the development of SUMI (the Software Usability Measurement Inventory; [9]). The developers of SUMI are also using a similar approach to develop questionnaire measures to assess the usability of web sites (WAMMI) and multimedia software (MUMMS). However, these measures are not claimed to be applicable to speech recognition systems.

In the field of speech-based systems research, many researchers have tried to measure subjective satisfaction, but none have invested any significant effort into the design and testing of their measuring instruments (see below). The research proposed here aims to fill this gap.

2.3. Measures of User Satisfaction with Speech Systems

A review of the literature on evaluation of speech systems shows that, while extensive consideration has been given to the definition and measurement of efficiency and effectiveness in user interactions with speech systems, comparatively little emphasis has been given to measurement of subjective satisfaction. For instance [5] authoritative handbook of standards for spoken language systems describes a large number of different efficiency and effectiveness measures for use in evaluation. In contrast their coverage of subjective satisfaction measures is fairly sparse, covering only the concept of using 5 or 7-point scales to obtain the judgement of test subjects on system properties. This limited coverage does not, however, appear to reflect a perceived lack of significance of user satisfaction, since the authors state that "[subjective assessment measures] can be very important for the global evaluation of a service or product, because in the end a human being has to use the system and if it is annoying or impractical it is likely that the system will be neither bought nor used".

A US-based research effort is currently developing the ‘PARADISE’ methodology for evaluating spoken dialogue agents [12]. Interestingly subjective satisfaction is placed at the centre of this model, showing the value which the authors place on this variable. The model aims to use decision theory to predict user satisfaction scores from objective measures of effectiveness and dialogue cost. However, the papers describing this work fail to describe how user satisfaction is measured in their studies. Given the centrality of user satisfaction in establishing the validity of the model, this would appear to be a serious limitation in their approach.

A number of individual research papers include mention of the use of one-off scales for subjective evaluation of speech systems ([2][13][3][10][8][4]). [2] for example, developed a scale of twelve bipolar adjective pairs (e.g. fast / slow) which were used as the end-points for 7-point rating scales. An overall “acceptability index” was calculated as the sum of the scores for each scale. [13] used a questionnaire format consisting of attitude statements (such as "The equipment is confusing to use") which were rated on a 7-point scale. [8] used a combination of attitude items to be rated on a 5-point scale and items with yes/no/maybe options. An overall "cumulative satisfaction" score was calculated from these items.

All of these one-off measures of user satisfaction, with minor exceptions, suffer from essentially the same weaknesses, most importantly:

• The techniques have not been satisfactorily validated, either against alternative subjective measures or against objective measures. Essentially the researchers have not established what it is that they are actually measuring with their instrument. Ideally the measures should be validated against real world behaviour, such that scores can be used to predict uptake and acceptance of a new system.

• There are no reports of the reliability of the techniques used. One issue is whether the results of the measures are stable over time (test-retest reliability). A second issue is the ‘internal’ reliability or consistency of a measure. This issue arises when multiple questions are used, each intended to contribute to the overall measurement of a ‘construct’ such as user satisfaction. This is a common approach in the measures reviewed, and is good practice because it reduces effects due to variations in questionnaire item interpretation. However, if this approach is used it is necessarily to provide evidence that all of the items really are measuring the same construct, otherwise the overall score will be meaningless.
None of the existing techniques for subjective speech interface evaluation reviewed meet the criteria for a valid (psychometric) measuring instrument. Unfortunately this means that the claims made on the basis of these measures (for instance, that a design parameter does or does not affect users attitude) must be treated with a great deal of caution.

2.4. Conclusion

Well designed measures of user satisfaction should make a significant contribution to the design of better speech-based interfaces. In order to overcome the limitations identified in existing measures we have begun a program of research with the aim of developing a validated and reliable measure for the Subjective Assessment of Speech System Interfaces (SASSI).

3. SASSI

3.1. The Questionnaire Development Process

We are using an empirical approach to questionnaire development [9], derived from work in the field of psychometrics. The process begins with the design of a large pool of attitude statements (such as "I enjoyed using this system"), with the aim of covering all aspects of the domain. This set of questions is used as the basis for an initial questionnaire, where people are asked to rate their agreement with each statement having just used a speech system. Once the questionnaire has been used by a large sample of people, using a range of speech products, the results are factor analysed in order to determine whether there is any underlying structure to the pattern of responses. If a structure is discovered, this indicates that the questionnaire is not measuring a uni-dimensional construct and instead is made up of various sub-scales, each measuring a different facet of user subjective experience with a speech system.

The questionnaire must then go through at least one more iterative cycle of design in order to confirm the suggested factor structure with a new set of data. Often this stage is also used to shorten the questionnaire (only retaining those items which load most strongly onto the proposed factors). Once this stage is completed the questionnaire must be evaluated for validity and reliability.

3.2. Work To Date

We have completed the first cycle of the iterative design process needed to create the SASSI measure. An initial 50 item questionnaire was designed after a review of the literature and interviews with speech system developers. Each attitude statement was rated according to a seven point scale labelled strongly agree, agree, agree, slightly agree, neutral, slightly disagree, disagree and strongly disagree. The measure was then used in a series of studies, evaluating eight different speech applications intended to represent the main styles of speech interface. Data was collected from 226 completed questionnaires and subjected to a factor analysis. Details of the analysis process are reported in [6].

The results suggest six factors in user attitude to speech systems which we have tentatively labeled: perceived system response accuracy, likeability, cognitive demand, annoyance, habitability and speed. For illustration the items loading onto the likeability factor are shown in Table 1.

3.2.1. Sub-scale reliability

The internal consistency of the items used to interpret each of the factors identified can be calculated using Cronbach's alpha. This is a means of assessing the extent to which items on a single sub-scale are measuring the same construct. Reliabilities of 0.80 or greater are generally required for widely used scales [7]. The sub-scale reliabilities that we obtained were: perceived system response accuracy, alpha=0.90; likeability, alpha=0.91; cognitive demand, alpha=0.88; annoyance, alpha=0.77; habitability, alpha=0.75 and speed; alpha=0.69. This indicates that three of our six sub-scales show a high enough internal consistency to merit general use in evaluation. Further work is needed to establish the validity of the remaining sub-scales.

Table 1: Items and Factor Loadings for Factor 2

<table>
<thead>
<tr>
<th>Item wording</th>
<th>Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system is useful</td>
<td>0.698</td>
</tr>
<tr>
<td>The system is pleasant</td>
<td>0.668</td>
</tr>
<tr>
<td>The system is friendly</td>
<td>0.621</td>
</tr>
<tr>
<td>I was able to recover easily from errors</td>
<td>0.606</td>
</tr>
<tr>
<td>I enjoyed using the system</td>
<td>0.587</td>
</tr>
<tr>
<td>It is clear how to speak to the system</td>
<td>0.578</td>
</tr>
<tr>
<td>It is easy to learn how to use the system</td>
<td>0.569</td>
</tr>
<tr>
<td>I would use this system</td>
<td>0.538</td>
</tr>
<tr>
<td>I felt in control of the interaction with the system</td>
<td>0.482</td>
</tr>
</tbody>
</table>

3.2.2. Correlation of sub-scales with a General Measure of System Quality

Users of the SASSI measure were also asked separately to give an overall rating to the system that they used, using the question "Overall I think this is a good system". Correlations between this measure and each of the sub-scales was as follows: Recognition performance, r=0.331; Likeability,
r=0.607; Cognitive demand, r=0.316; Frustration, r=0.386; Habitability, r=0.033; Speed, r=0.200. All of these correlations (expect Habitability) were found to be significant (p<0.005).

While the conclusions which can be made from comparison against a single measure of this kind are clearly limited, it is interesting to note that the highest correlation was with the Likeability sub-scale. This suggests that Likeability is the construct which is closest in meaning to users’ very general subjective opinions as to what is a good speech system.

4. Discussion

The research summarised here suggests six main factors which contribute to a user's subjective experience with speech-based systems. We have tentatively named these perceived system response accuracy, likeability, cognitive demand, annoyance, habitability and speed. The reliabilities of the sub-scales suggest that all are worthy of further research effort. The reliabilities of the perceived system response accuracy, likeability, cognitive demand sub-scales appear appropriate for general use, though of course these values will need to be confirmed with a new sample of data, and further research is planned to do this.

Further work is also needed to investigate the validity of the SASSI measure. For instance we plan to conduct research to establish whether SASSI scores are able to predict user preferences among alternative products. If this is shown to be the case then the measure will have great potential for use in the development of speech-based products for the future.

The statistical procedures needed to validate the questionnaire rely on large samples of data, so the biggest challenge we face is in collecting sufficient numbers of completed questionnaires. We aim to encourage as many researchers as possible within the speech research community to use our prototype questionnaire to enable this future development work.

5. Conclusions

While we have not yet completed the development of the SASSI measure, three of the sub-scales which we have investigated so far have an advantage over previous measures used in this area in that they have been shown to have acceptable values for internal consistency. The likeability sub-scale, as reproduced in Table 1, in particular has the highest Cronbach's Alpha value and correlates most strongly with a general measure of speech satisfaction. Until further research on SASSI is completed we would advise the use of this sub-scale by researchers who need a measure of subjective satisfaction with speech systems.

6. References


