Usability measurement for speech systems: SASSI revisited

Kate Hone
Brunel University
Uxbridge, UB8 3PH, UK
kate.hone@brunel.ac.uk

ABSTRACT

In 2000 Hone and Graham [4] published 'Towards a tool for the subjective assessment of speech system interfaces (SASSI)'. This position paper argues that the time is right to turn the theoretical foundations established in this earlier paper into a fully validated and score-able real world tool which can be applied to the usability measurement of current speech based systems. We call for a collaborative effort to refine the current question set and then collect and share sufficient data using the revised tool to allow establishment of its psychometric properties as a valid and reliable measure of speech system usability.

Author Keywords

Speech recognition, spoken dialogue systems, usability.

ACM Classification Keywords

Human factors.

INTRODUCTION

The ability to accurately measure speech system usability is a vital component for both good system design and theoretical research on designing speech and language interactions. Without good measurement we cannot assess whether systems have been improved through redesign, nor can we evaluate which particular design features affect usability, nor can we empirically compare alternative system designs. 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 the late 1990s we made some initial progress towards the development of a tool to measure the subjective satisfaction of speech system interfaces (SASSI) [4]. The SASSI questionnaire was originally developed as a collaboration between two UK funded speech projects. An initial set of questionnaire items was generated though reviewing the literature at the time and consulting with experts. This questionnaire was then piloted with a sample of over 200 respondents and the results were factor analyzed to uncover the underlying factor structure. The results led us to propose fix factors in user attitudes to speech systems: perceived system response accuracy, likeability, cognitive demand, annoyance, habitability[1] and speed. These results were potentially important as they revealed the salience of a number of features which are not typically found in the subjective attitudes of users of more traditional types of interface (and reflected in standard usability questionnaires). This suggests the need for the development of specific measures of usability for speech systems. The development work for SASSI included only the first stage of the type of iterative development which is needed to establish the validity and reliability of a measurement instrument. Further work is needed to refine the instrument and establish its psychometric properties.

This position paper is organized as follows. We begin by justifying the importance of subjective satisfaction for speech recognition systems. We then describe approaches to the measurement of user satisfaction in general before describing the particular case of measuring user satisfaction with speech based systems. We then lay out an agenda of the further work which is needed to take SASSI forward as general tool which can be applied in this domain. We conclude by calling for collaborators for a joint effort to further develop the tool.

THE IMPORTANCE OF SUBJECTIVE SATISFACTION WITH SPEECH SYSTEMS

We would argue that 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

[1] Habitability refers to whether there is is a good match between the user's conceptual model of the system and the actual system behavior [3]

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 nature of recognition technology is such that there will always be occasional failures in recognition. Given this feature, it is therefore vital that we know what level of performance users will tolerate in which kinds of context. Objective measures of system performance, though important, are not sufficient.

MEASURING USER SATISFACTION

When we measure something there are certain qualities we require from our measuring instruments, 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 [10]:

- 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, for instance in the physical sciences you would not design a measuring tape made out of elastic (reliability) or measure the diameter of an atom with a meter rule (sensitivity). 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), 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) [8]. However, these measures are not claimed to be applicable to speech recognition systems and our pilot work in SASSI suggests that they are not wholly appropriate in this domain.

While SASSI represented a move in the right direction, in the field of speech-based systems research, the required degree of effort has not yet been invested into the design and testing of subjective measuring instruments.

MEASURES OF USER SATISFACTION WITH SPEECH SYSTEMS

The importance of subjective assessment of speech systems is well recognised. For example [2]'s handbook of standards for spoken language systems states 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". In the 'PARADISE' methodology for evaluating spoken dialogue agents [7] subjective satisfaction is placed at the centre of this model, showing the value which the authors place on this variable. However, despite the importance placed on the idea of speech system usability, progress towards a validated test instrument is limited.

The most notable progress in the right direction can be seen in the work of [9] who attempted to measure the reliability and validity of a modified version of SASSI and the questionnaire from list of questions are proposed in ITU-T Rec. P.851 [6]. While their results were promising in supporting the usefulness of both questionnaire measures, the sample sizes used in this study was really too small to justify the use of factor analysis.

AGENDA FOR FUTURE DEVELOPMENT OF SASSI

Aims

We argue that further work is needed in order to develop a measure of user satisfaction for evaluating systems which use speech in their interfaces (either for input using speech recognition, output using recorded or synthesised speech, or both).

The specific research objectives are to produce a subjective tool which is:

1. Valid, reliable, sensitive and free from contamination.
2. Widely applicable to all styles of speech interface (for instance from command and control to natural language).
3. Quickly and easily completed by naïve and/or first time respondents.
4. Quantifiable, to allow statistical comparison of multiple design alternatives, or benchmarking of a single product during development.
5. Complete, capturing all important aspects of a user's experience with a speech system.

Methodology

We propose that an empirical approach to questionnaire should be adopted in order to take SASSI forward. The empirical approach to questionnaire design begins with the production of a large pool of questionnaire items, intended to sample all attitudes relevant to the domain. Empirical methods are used to determine latent structure from the pattern of responses to these questions. The questionnaire is then refined based on what is learnt and the process is begun again. Once a structure is established and confirmed, steps must be taken to establish the validity, reliability and sensitivity of the scale or scales produced. The inital work to produce SASSI means we will be beginning the second cycle of iterative development.

Research steps

Step 1: Pilot Questionnaire Development

The current SASSI questionnaire will form the starting point, but several important alterations will be made to it. First, questionnaire items will be added to address factors which were proposed in SASSI, but were not adequately covered by the current question set (e.g. speed of interaction). We will also draw on the research such as [9] which has taken place since SASSI was originally published and consult with experts in speech & language to ensure adequate item coverage. Second, the scope of the measure will be extended to allow measurement of subjective responses to speech output.

Step 2: Data collection and analysis

During this phase the pilot questionnaire will be made available to the community and we would encourage researchers and companies to use it in their ongoing product / service evaluations, allowing us to have access to anonymised responses. During this phase we would hope to collect questionnaire data from at least 300 users of a range of speech-based applications to allow factor analysis. We would then investigate whether the results support the initial factor structure suggested in SASSI; we would also investigate the factor structure of the new speech output component.

The questionnaire would then be redesigned such that a smaller number of items mapping onto each factor would be retained. The aim will be to produce a shorter questionnaire which can be ***quickly and easily completed***.

Step 3:Iterative testing and analysis .

Steps 1 and 2 would need to be repeated for the new version of the questionnaire.

Step 4:Validity evaluation

The aim during this phase would be to collect SASSI measures for live speech products and compare responses to other metrics of system success (such as sales, usage data, completed interactions, etc). Correlation/regression

would provide a measure of validity. SASSI evaluation of near to market products, followed up by correlation/regression with subsequent success could provide a measure of ***predictive validity***.

Step 5: Reliability and Sensitivity Evaluation

Further work would be needed to evaluate the ***test-retest reliability*** of the measure and the sub-scale ***reliability*** (internal consistency).

Test-retest reliability could be investigated by asking a group of participants to use a speech system and rate it using the SASSI measure, and then repeat the process after a delay of approximately two months. The degree of correlation between the two sets of data indicates the test-retest reliability.

Although estimates of the internal reliability / consistency of the sub-scales can be obtained when the initial factor analysis is performed, this should be confirmed with an independent sample of data. During this phase of the research we would again need to encourage companies to use the final SASSI measure in their on-going product evaluations. We would aim to collect at least 300 completed questionnaires from users of a range of speech applications. This data would then be used to calculate Cronbach's Alpha values for each of the individual sub-scales making up SASSI. Ideally we hope that all values will be at least 0.80, the level generally required of widely used scales [5].

A program of experimental research would be needed to evaluate the ***sensitivity*** of the measure.

Step 6: Questionnaire release

The ultimate aim would be to allow the release of the questionnaire measure with associated documentation for use by the community at large. This documentation is important as it will indicate how the measures are to be interpreted. Relevant documentation might include by-country norms for the measure (i.e. average score for each sub-scale based on the assessments of speech products). This would allow those who wish to use SASSI to evaluate a single product to tell whether their product is better than, or worse than, the current average. Overall the documentation will support the aim of developing a ***quantifiable*** measure, allowing statistical comparison of alternatives and benchmarking.

It is envisaged that the SASSI measure will be made freely available to academic researchers to allow further research in this area.

CONCLUSION

The key points of this paper are that:

i) Speech interfaces will play a major role in future interactive systems.

ii) Well designed measures of user satisfaction will make a significant contribution to the design of better speech-based interfaces.

iii) There is therefore a need for validated and reliable measures of subjective satisfaction when using speech systems.

We hope that cross-community collaboration on this issue could bring us closer to this goal.

REFERENCES

1. Bevan, N. (1994) Ergonomic Requirements for Office Work With VDT's, Technical Report 9241-11, ISO.
2. 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.
3. Hone, K.S. and Baber, C. (2001) Designing Habitable Dialogues for Speech-based Interaction with Computers. International Journal of Human Computer Studies, 54(4), 637-662.
4. Hone, K.S. and Graham, R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering, 6(3/4), 287-305.
5. Igbaria, M. and Parasuraman, S. (1991) Attitudes towards microcomputers: development and construct validation of a measure. International Journal of Man-Machine Studies, 35, 553-573.
6. ITU -T Rec. P.851, 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.
7. 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.
8. 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.
9. Möller, S., Smeele, P., Boland, H., Jan Krebber, J. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study, Computer Speech & Language, Volume 21(1), 26-53.
10. Sanders, M. S., & McCormick, E. J. 1993. *Human Factors in Engineering and Design*. (7th ed.). New York: McGraw-Hill.