UNDERSTANDING SPEECH UNDERSTANDING

R. K. Moore

DERA Speech Research Unit
St. Andrews Road, Malvern
Worcs., WR14 3PS, UK

ABSTRACT

Despite the significant theoretical and practical advances that have been made in automatic speech recognition in recent years, relatively little effort has been devoted to the evaluation of speech in an interactive multi-modal application interface. This paper introduces a general methodology for assessing speech-based systems and concludes with a proposal for a test scenario which focuses on the understanding component of a spoken language system.

1. INTRODUCTION

Recent years have seen a substantial growth in the capabilities of automatic speech recognition (ASR) both in the research laboratory and in the commercial marketplace [3][7]. In something over a decade, the technology has developed to the point where very large vocabulary speaker-independent continuous speech recognition (LVCSR) is available 'off-the-shelf' for only a few tens of dollars.

This steady improvement in capability has been fuelled by a number of developments: the relentless increase in desktop computing power, the introduction of hidden Markov modelling (HMM), and the existence of the annual round of LVCSR system evaluations sponsored by the US Defence Advanced Research Projects Agency (DARPA) programme.

However, the recognition of spoken utterances is only one aspect of a speech-enabled interface. Whilst the high-profile research developments (and mainstream commercial offerings) have been targeted at the transcription of dictated documents, a considerable body of equally important research and development has been focused on the use of speech for eyes-free/hands-free control of application functions or remote access to information. In such situations the speech channel sits alongside other important input/output modalities such as keying, pointing, imaging and graphics. This means that the actions and behaviours of the speech-specific components of a spoken language system have to be carefully orchestrated with respect to other modalities and the analysis of a spoken utterance often has to go beyond a straight transcription of what has been said. An interpretation of the semantic 'intent' of the user is required – in other words, there has to be some degree of 'understanding' [26][6] [1] [9][5].

Also, in recent years, there has been a considerable amount of attention devoted to 'dialogue systems' [20][23] in which real-time interaction between a user and an application is a key requirement. As illustrated in Fig. 1, the dialogue component of a spoken language system is seen as central to the interface between the linguistic and spatial interpreter-generator input-output modes and the application itself.

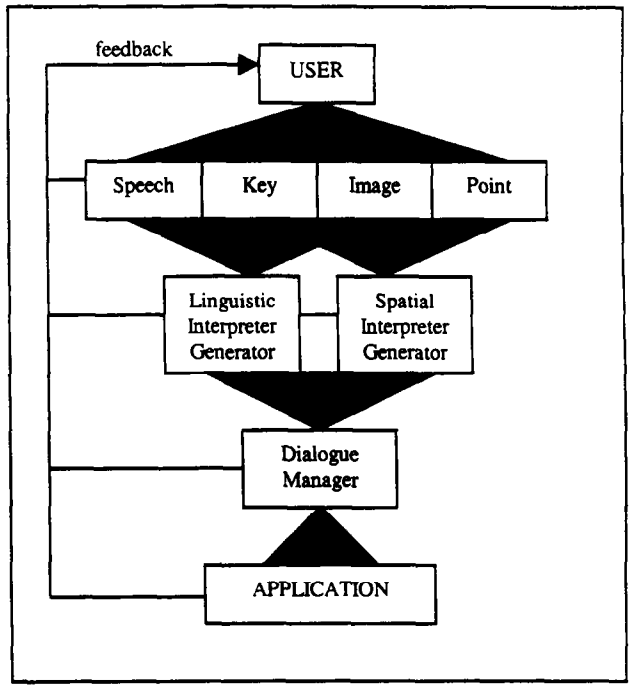

Figure 1. Multimodal human-computer interface.

Clearly, a productive dialogue could not take place between a user and an application without an appropriate level of interpretation of the meaning of different spoken inputs. However, much of the work so far has been directed towards highly constrained tasks with prescribed dialogue structures and correspondingly limited syntactic and semantic representational frameworks.

One common feature shared by work in speech understanding and that on spoken dialogue systems is the lack of clear performance metrics. As yet there is no agreed definition of what constitutes a 'good' dialogue (beyond the superficial notion of minimum transaction time) - as has been said in another context, "it is not that it works well that is of interest, it is the fact that it works at all". Of course it is easily possible to speculate on the many and varied features of an effective speech-based interface, but this is the realm of human factors and, as such, the speech field has yet to draw significantly on the relevant body of expertise.

Performance measures for understanding, on the other hand, are even less well developed [2]. Proposals have been made and issues have been discussed, but the general conclusion seems to be that there is a need to agree on a common framework for semantic-level representations [13] and, because of the theory-dependent nature of alternative formalisms, such agreement is difficult to achieve.

What is required is a complete re-appraisal of the evaluation framework for speech-based systems. Clearly, there is a need to be able to assess individual system components, but perhaps it is now time for this to be done in the context of complete working systems.

2. A GENERAL METHODOLOGY FOR ASSESSMENT

One possible model for a general assessment framework for spoken language systems is illustrated in Fig. 2 [17]. A key feature of this model is its accommodation of the complex relationship between the technical features and the operational benefits of spoken language systems. The model also makes clear that a meaningful definition of the general 'suitability' of a given technology for a particular application is dependent on a multi-factorial assessment along both technical and operational dimensions – the matching of relevant requirement and capability profiles.

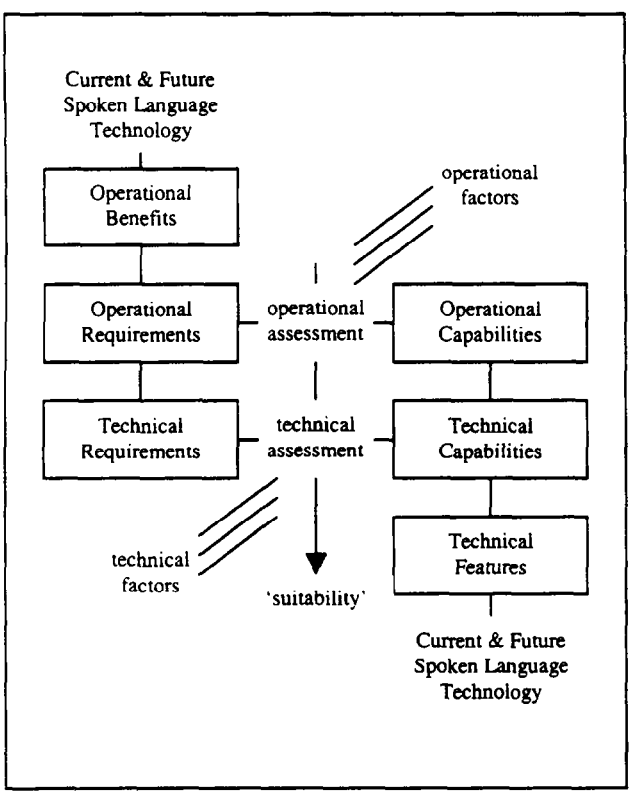

Figure 2. A model of the relationship between the applications of spoken language systems and the underlying technology.

What this points to is the possibility of establishing a comprehensive 'whole-system' approach to evaluation driven by the need to deliver clear operational benefit. All components of a system influence the derived benefit to a greater or lesser extent and, in order to characterise this, a weighted account should be taken of all influencing variables. Notions of a 'correct' recognition, a 'good' dialogue or an 'appropriate' interpretation become subservient to the idea of an 'optimal' operational benefit.

3. WHOLE-SYSTEM ASSESSMENT

In many applications, the key operational benefits are:

- Cost savings
- Manpower savings
- Increased operational effectiveness (e.g. in defence applications)
- Increased productivity (e.g. in civil applications)
- Workload reduction
- Increased security
- Increased safety
- Increased functionality
- Space savings
- Bandwidth savings
- Improved quality of life

In the same way that the technical features of an automatic speech recogniser (e.g. whether or not it uses HMMs) have a complex relationship with it's technical capabilities (e.g. the vocabulary size), so the operational benefits enjoy a complex relationship with the operational requirements. For example, the cost savings that might be possible by the introduction of spoken language technology depend on a range of operational factors such as accessibility, interoperability, portability, mobility, re-configurability, etc. However, for a whole-system assessment, performance *must* be measured along each of these dimensions.

In this way the 'suitability' of spoken language technology in general, or of a particular technical offering, can be judged against competing solutions.

Unfortunately, the area which is least well understood is the critical relationship between the operational requirements & capabilities and the technical requirements & capabilities. For example, at some point it is necessary to be able to translate an operational requirement such as overall 'efficiency' into a technical requirement such as recognition 'word error rate'. However, this will become better understood as more systems enter the marketplace, and studies are undertaken to understand these key characteristics of genuine applications. The payoff will be an ability to field more *appropriate* solutions thereby addressing the concerns of professional 'computer-sceptics' [4].

4. TESTING UNDERSTANDING

From the forgoing, it is clear that 'understanding' can be tested at the technical or operational levels. At the technical level – that is, the level of a system component – problems arise due to the difficulty of obtaining agreement on defining suitable metrics (as discussed in Section 1). At the operational level, on the other hand, it is only necessary to focus on measures of general

effectiveness, and this could open up more possibilities for defining acceptable (non theory-dependent) test scenarios.

Interestingly, this line of argument goes back to the days of the original ARPA Speech Understanding programme of the early 1970s [10]. At that time, there was considerable concern that high-accuracy acoustic-based ASR would be impossible without the strong top-down involvement of higher-level constraints (such as those imposed by the application and the language) [18]. As a consequence, contemporary speech understanding systems (SUS) focused on the requirement that they should perform the correct actions, rather than recognise each word correctly.

4.1 On-Line vs. Off-Line Testing

One important issue is whether understanding ability can (or should) be tested in a non-interactive situation. The recent DARPA-funded work aimed at an air travel information service (ATIS) [19] was obliged to address this issue. Even though the scenarios were established using interactive 'Wizard of Oz' (WOZ) techniques, the testing had to resort to a categorisation of individual utterances in terms of their independence from each other. This allowed the initial research to focus on 'one-shot' single-input single-output processes.

At the other extreme, the staging of the ELSNET-ELRA 'Olympics' on "Testing Spoken Dialogue Information Systems Over The Telephone" at Eurospeech'97 assessed complete transactions. Ten live systems competed, and were judged by a large number of conference participants according to a range of *subjective* criteria.

Of course, the difficulties associated with conducting *objective* interactive assessments are manifest. First, there are the linguistic interdependencies between utterances (mentioned above). Second, there is the difference between transactional and clarification dialogue. The third problem area is the dynamic nature of a dialogue (as a function of changed system variables). The final problem is the possible adaptive (i.e. context-dependent) behaviour of a whole range of system components.

The challenge is to come up with an assessment framework which is able to finesse all these difficulties.

4.2 Prior Art in Text Understanding

The field of natural language processing (NLP) has been grappling with these issues for some time, most notably within the DARPA-sponsored 'text retrieval' (TREC) and 'message understanding' (MUC) conferences. The general approach is to view the process as one of 'form filling' and, much as in a general linguistic comprehension test, the requirement is to determine "who did what to whom, and why".

Clearly this 'template-based' approach is somewhat similar to the construction of database queries in ATIS. In this case, the essential information relating to destinations and starting times etc. needs to be extracted from the speech and entered in the appropriate fields of a query.

A completely different, but whole-system, approach is to apply the 'Turing test' [25] i.e. to determine to what extent is it possible to convince a user that an automated system is human-like. Somewhat along the lines of the Eurospeech Olympics, the NLP community enjoys an annual public event at which the 'Loebner prize' is awarded to the best such system. Although this approach has its critics [22], it would be interesting to debate the feasibility of a speech-based entry (especially since speech, being so much more expressive than text [8], would bring the event much closer to Turing's original ideas).

4.3 Towards A Solution

The main problem with the Loebner prize, and to some extent the Eurospeech Olympics, is the subjective nature of the assessments. It might be relatively easy to fool a user into judging that an interaction with an automated system is 'natural' if there is no other purpose than to chat [24]. However, whilst it may be very important to develop a technology which is capable of chatting [11], the social consequences (beneficial or otherwise) may be difficult to quantify.

A worthwhile solution needs to rest on *objective* measures; it must be possible to establish the 'ground truth' and, ideally, this should be *theory-independent*. It is also important to encompass *whole-systems* and *complete tasks*, and to invoke measures which relate to *operational benefit* and which are functions of a profile of *dependent variables*. It should also be open to WOZ-style simulation.

5. A PROPOSAL

A solution which fits all of the foregoing criteria is one in which a user describes a visual scene using their voice. The task of the automated system would thus be to re-construct a given visual scene within a synthetic environment.

In this scenario an objective comparison can be made between the original and the re-constructed (interpreted) scene and, in essence, such a task can be viewed as a process of 'speech-to-image translation' (as envisaged over twenty years ago [15][16]). Whether the process is one-shot or incremental, the key notion is that the end result is a direct indication of the degree of 'understanding'. Therefore, in this task, understanding capability can be both measured and calibrated with respect to a wide range of dependent variables (such as transaction time, word error rate etc.) – these being the key research challenges.

One advantage of such an approach is that the difficulty of the task could be scaled in a controllable manner. For example, a simple scene might consist of a set of abstract coloured (3D) shapes, and the task would involve the expression of straightforward spatial relationships and object descriptions. At the other end of the scale, a complex scene might involve a natural image (such as a photograph), and the task could be concerned with the complex physical and spatial inter-relationships between a wide range of information-bearing elements and forms.

Another advantage of this proposal is that it has a natural analogue in the defence arena – 'reconnaissance reporting' - a task already studied in an LVCSR context [21]. As such, it is likely to be of interest to the funding agencies traditionally associated with pushing forward the performance envelope of spoken language systems.

The proposal also focuses on a task which does not need to be interactive verbally. In other words, it is essentially a mixed-mode application – speech-in image-out – and this removes the requirement for clarification dialogue and natural language generation.

Interesting issues which arise from this basic framework include:

- The implementation could be interactive or non-interactive (the first requiring near real-time recognition and understanding).
- A requirement to handle 'out of world' (OOW) objects – analogous to, but much more interesting than, 'out of vocabulary' (OOV) words.
- The need to define a *spatial* metric for judging interpretation accuracy.
- The option of performing WOZ studies to facilitate comparisons with human performance [14][12].
- A requirement to 'calibrate' a user's ability to be understandable.

6. CONCLUSION

This paper has argued that it is now appropriate to move towards a general methodology for the whole-system assessment of multi-modal speech-based systems. In particular, it has been suggested that there should be a focus on the operational benefits to be derived in any given application, and that these need to be mapped onto the relevant operational and technical requirements. The testing of the understanding capabilities of a spoken language system has been discussed and a proposal has been presented for a 'speech-to-image translation' (verbal scene description) task. It is recommended that serious consideration be given to the opportunities presented by this scenario.

7. REFERENCES

[1] Bates M., Bobrow R., Fung P., Ingria R., Kubala F., Makhoul J., Nguyen L., Schwartz R. and Stallard D. "The BBN/HARC spoken language understanding system". *IEEE International Conference on Acoustics Speech and Signal Processing*, vol.II, pages 111-114, 1993.

[2] Cole R., Hirschman L., Atlas L., Beckman M., Biermann A., Bush M., Clements M., Cohen J., Garcia O., Hanson B., Hermansky H., Levinson S., McKeown K., Morgan N., Novick D., Ostendorf M., Oviatt S., Price P., Silverman H., Spitz J., Waibel A., Weinstein C., Zahorian S. and Zue V. "The challenge of spoken language systems: research directions for the nineties". *IEEE Transactions on Speech and Audio Processing*, vol.3, pages 1-21, 1995.

[3] Comeau P. "Voice control". *PC Plus*, pages 222-229, April 1997.

[4] Gibbs W. "Taking computers to task". *Scientific American*, pages 82-89, July 1997.

[5] Glass J., Flammia G., Goodine D., Phillips M., Polfroni J., Sakai S., Seneff S. and Zue V. "Multi-lingual spoken-language understanding in the MIT Voyager system". *Speech Communication*, vol. 17, pages 1-18, 1995.

[6] Hoge H. "SPICOS II – A speech understanding dialogue system". *International Conference on Spoken Language Processing*, vol.2, pages 1313-1316, 1990.

[7] Honeyball J. "Power of speech". *PC Pro*, pages 255-257, September 1997.

[8] Hunt M. "Speech is more than just an audible version of text". *The Structure of Multimodal Dialogue*, M. Taylor, F. Neel and D Bouwhuis (eds.), North-Holland, 1991.

[9] Issar S. and Ward W. "CMU's robust spoken language understanding system". *Proceedings of Eurospeech Conference*, vol.3, pages 2147-2150, 1993.

[10] Klatt D. "Review of the ARPA speech understanding project". *Journal of the Acoustical Society of America*, vol.62, pages 1345-1366, 1977.

[11] Locke J. "More than words can say". *New Scientist*, pages 30-33, 18 March 1995.

[12] Lippmann R. "Speech recognition by machines and humans". *Speech Communication*, vol.22, pages 1-16, 1997.

[13] Moore R. C. "Semantic evaluation for spoken-language systems". *Proceedings of the Human Language Technology Workshop*, Morgan Kaufmann, 1994.

[14] Moore R. K. "Evaluating speech recognizers". *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.25, pages 178-183, 1973.

[15] Moore R. K. "A descriptive technique for the analysis and design of speech understanding systems". *PhD Thesis*, University of Essex, 1976.

[16] Moore R. K. "A multilevel approach to pattern processing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.1, pages 86-88, 1979.

[17] Moore R. "Users guide". *EAGLES Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore and R. Winsky (Eds.), Mouton de Gruyter, pages 1-28, 1997.

[18] Newell A., Barnett J., Forgie J., Green C., Klatt D., Licklider J., Munson J., Reddy R., and Woods W. *Speech Understanding Systems*. North-Holland/American Elsevier, 1973.

[19] Pallett D., Fiscus J., Fisher W., Garofolo J., Lund B., Martin A. and Przybocki A. "1994 benchmark tests for the ARPA spoken language programme". *Proceedings of the Spoken Language Systems Technology Workshop*, Morgan Kaufmann, pages 5-36, 1995.

[20] Peckham J. "Speech understanding and dialogue over the telephone: an overview of progress in the SUNDIAL project". *Proceedings of Eurospeech Conference*, vol.3, pages 1469-1472, 1991.

[21] Russell M., Ponting K., Peeling S., Browning S., Bridle J. and Moore R. "The ARM continuous speech recognition system". *IEEE International Conference on Acoustics Speech and Signal Processing*, 1990.

[22] Scheiber S. "Lessons from a restricted Turing test". *Communications of the Association for Computing Machinery*, 1994.

[23] Sutton S., Novick D., Cole R., Vermeulen P., de Villiers J., Schalkwyk J. and Fanty M. "Building 10,000 spoken dialogue systems". *International Conference on Spoken Language Processing*, vol.2, pages 709-712, 1996.

[24] Suzuki N., Inokuchi S., Ishii K. and Okada M. "Chatting with interactive agents". *Proceedings of Eurospeech Conference*, vol.4, pages 2243-2246, 1997.

[25] Turing A. "Computing machinery and intelligence". *Mind*, LIX(236), pages 433-460, 1950.

[26] Young S. "Use of dialogue, pragmatics and semantics to enhance speech recognition". *Speech Communication*, vol.9, pages 551-564, 1990.