

## THE PREDICTION OF SPEECH RECOGNISER PERFORMANCE BY THE USE OF LABORATORY EXPERIMENTS: SOME PRELIMINARY OBSERVATIONS

Trevor J. Thomas, Logica Cambridge Ltd, 104 Hills Road, Cambridge, CB2 1LQ

### ABSTRACT

Speech recogniser assessment, an area of speech recognition research which is generating a great deal of interest. Currently available assessment techniques tend to produce inaccurate or unreliable results, this has been seen to be due to essential limitations in the methodologies adopted.

This paper describes an assessment methodology designed to overcome many of the limitations of the current techniques. It has a mathematical basis and uses a statistically derived data base to develop standard tests for various environments of interest. It is hoped that by the use of these techniques experimenters will be able to perform laboratory based recogniser tests with confidence. It is expected that for many applications these laboratory tests will not replace field trials, but will provide guidance towards suitable recognisers.

### INTRODUCTION

Speech recogniser performance assessment is a field of speech recognition research which is currently being investigated by many researchers throughout the world. Their intention is to find some techniques that can reliably be used to measure recogniser performance in some standard fashion. This paper examines the problems that are faced, and the solution that is being investigated by Logica Cambridge.

Currently there are two main ways in which recogniser assessment can be performed, either field trials can be performed, or pre-recorded speech data bases containing a representative selection of likely speech tokens can be used. Both methods use the recogniser on test to directly process the incoming speech, and both methods will provide a single number which will be a measure of the recognisers performance. Field trials are probably the more accurate of the two approaches, but they are expensive to perform and elaborate to control. Pre-recorded data bases, once recorded, are simpler to use, but may be less representative of field conditions. Pre-recorded data bases also constrain the investigator in his choice of experiments, the only speech available to him would be that in the data base and he would not be able to examine the effects of any other type of speech. We would like to present, in this paper, an approach to recogniser assessment which uses a data base with a different, and controlled, structure coupled with a mathematically based methodology which would be suitable for more representative experimentation and therefore provide more accurate results.

### REQUIREMENTS FOR THE METHODOLOGY

To assess a recognisers performance, in any particular environment, we must have available a series of experiments which will be sufficient for an investigator to make a comprehensive statement on the recognisers capability.

Because speech recognition is a complex action, including not just the recogniser itself, but also the interaction between the speaker's environment, the speaker's own voice characteristics, the higher level processing undertaken by the driving computer and the type of feedback given to the speaker, it would be unreasonable to expect that a single performance figure could be produced which would be taken to be *the* recogniser's

performance. Figure 1 shows many of the conditions which could be expected to affect the recogniser. We hope that a series of experiments, each tailored to one specific aspect of the recognition procedure, would produce a set of results in which each separate result would describe a single aspect of the recogniser's performance, and which, when taken together, would produce a more meaningful overall picture. This is in many ways, analogous to the use of bench-marks in computer performance analysis. The overall performance differences between two computers cannot be gleaned by examining the MFLOPS figures alone, but should also include disk accessing times and other relevant figures. It is difficult to take the analogy further though, because of the techniques used to perform the assessment.

### RECOGNISER SENSITIVITY ANALYSIS

The method discussed here, Recogniser Sensitivity Analysis (RSA), is due to Knight and Peckham (1982). The essentials of this method require that a database be constructed with a carefully controlled range of variability in many speech parameters, it should then be possible to test a recogniser's sensitivity to many different parameters and from these predict its performance under field conditions using a manageable amount of data. The concepts upon which it is based allow experimenters to take 'standard' database subsets off the shelf and to use these to benchmark the performance of their recognisers. Such databases may contain tokens suitable for the evaluation of recognisers in environments such as high performance jets, offices and telephone systems, and they may contain various types of vocabulary, perhaps isolated digits, connected speech, or vocabularies chosen to be of particular interest to various manufacturers. They would also contain tokens of a wide range of speakers so that pre-speech production effects can be examined. It may also be possible for this 'benchmarking' to be done by the manufacturers and published with other information about their products.

Essentially, RSA includes the capability to perform application non-specific, vocabulary independent performance assessment though the use of a carefully constructed database accompanied by a mathematically based model of speech variability and recogniser performance.

There is a need to reduce the data requirements of RSA, and this may be achieved by using analysis of variance, and then possibly, Response Surface Methodology (RSM). To these could be added the entropy based Relative Information Loss metric (RIL) of Shannon and Weaver (1964) with enhancements for speech recognition by Moore (1977). This data reduction is not essential to the conceptual functioning of RSA, but rather relates to an implementational detail. The methodology can be separated into two sections, one is concerned with the analysis of the environment to prepare the system for testing, the other is the testing procedure itself, generally, only the second stage will be required by users. The first task of the methodology is to analyse the recogniser's environment, which includes the speaker, into a small number of features that can be extracted which can then be used to perform the experiments. This data reduction can be summarized as the production of a set of features derived from the environmental characteristics by a combination of the quantifiable parameters as recorded from the environment. To produce the 'standard' databases, analysis of variance techniques are being used to select speech tokens which have the correct values of the features. These speech tokens will comprise the 'standard' database and will then be available to users who wish to assess their recognisers in this way. The features that are being investigated are designed to encompass the entire range of speech production parameters that any recogniser is likely to meet. These parameters are summarized in Table 1.

---

Voice Quality - Breathy, Croaky etc.

Syllabic Rate

Mean Pitch

Vocal Tract Shape

Speaking Intensity

Phonetic Confusibility

Speech Intelligibility

---

### Table 1. Speech Production Parameters

---

All of these parameters are measurable either automatically, in the cases of syllabic rate, mean pitch, speaking intensity and speech intelligibility, or by simple measurements upon the speaker. The basis for choosing tokens which have the correct features is the Relative Information Loss metric. In this model, the speaker uses his speech to encode his intentions whilst the recogniser acts as the decoder, any ambiguity introduced, by any means, into the transmission (eg. background noise) is treated as added interference. Classically, the RIL is a statement of the entropy gain through a transmission network. If the vocabulary used by the speaker is  $V$  with each item represented by  $v_i$ , and the vocabulary contains  $N$  items, and the probability of  $v_i$  being uttered is  $P(v_i)$ , then the measure of information content in the transmission, the entropy  $H(V)$  is,

$$H(V) = - \sum_{i=1}^N P(v_i) \log_2(v_i)$$

A higher  $H(V)$  implies a lower redundancy and a harder recognition task. To determine the information loss; if the vocabulary of possible recogniser output symbols  $U$  is  $u_1, \dots, u_M$ , the probability receiving symbol  $u_i$  is  $P(u_i)$ , and the probability that  $v_i$  could be recognised as  $u_j$  is  $P(v_j, u_i)$ , then,

$$H(U,V) = \sum_{i=1}^N \sum_{j=1}^M P(u_j) P(v_i, u_j) \log_2 P(v_i, u_j)$$

The RIL is then,

$$RIL = \frac{H(U,V)}{H(V)}$$

If  $RIL = 1$ , the transmission system has perfect performance, the smaller the RIL, the worse the recogniser's performance. This metric is used as the distance metric throughout RSA, although other, equally useful metrics have been proposed (Taylor 1979).

## CONCLUSIONS

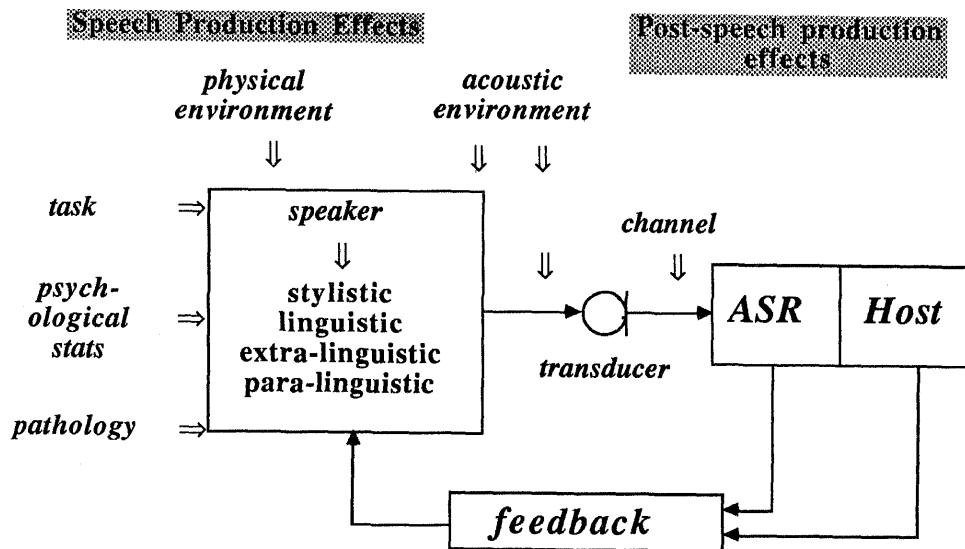
It is hoped that the methodology proposed in this paper, which includes carefully constructed databases, relational database management and mathematically based analyses will be capable of reliable speech recogniser assessment in a low-cost environment which would be both suitable for both recogniser algorithm development and product evaluation.

## ACKNOWLEDGEMENTS

This work is being funded by the Department of Trade and Industry as part of the Alvey programme. The national Physical Laboratory, Smith's Industries Ltd, University College London and The Royal Signals and Radar Establishment are the other members of the consortium.

## REFERENCES

- 1 J.A.Knight, J.B.Peckham. "A Generic Model for the Assessment of Speech Input Applications," Logica report for RSRE August 1984.
- 2 C.E Shannon, W. Weaver. "The Mathematical Theory of Communication," University of Illinois, 1964.
- 3 R. Moore. "Evaluating Speech Recognisers," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol ASSP-25, April 1977
- 4 M.M Taylor. "Issues in the Evaluation of Speech Recognition Systems," Draft of NATO AC243 Report May 1980
- 5 T. Frangoulis, Private communication, May 1987



**Figure 1. Environmental conditions affecting speech recogniser performance**