ABSTRACT

All manufacturers of speech recognition equipment claim near perfect recognition performance for their products, of 98% correct recognition or better. Few manufacturers put their performance rates into context, by describing the circumstances in which these measurements were made. In practice, performance rates in excess of 95% are achieved with very small vocabularies of dissimilar words, in ideal environments with consistent speakers.

In recent years, much research has been carried out by various groups to develop a test procedure and analysis technique to determine speech recogniser performance, but as yet none have been accepted as standard. Experimental design is critical if meaningful results are to be achieved.

This paper reports the findings of a research project investigating aspects of experimental design in speech recognition performance tests. A series of design problems are discussed including how to eliminate, or at least compensate for, all the factors, acoustic, environmental, dialectic, idiolectic, physical, psychological and technical, that can vary during a test and cause the recogniser to make a mistake. Methods of running tests are compared along with statistical procedures for analysing results.

Finally, the results of a series of voice tests are discussed. The tests were carried out with the aim of determining the influence of several factors on recognition performance. These include the sex, age and experience (or using voice recognition equipment). Subjects participating in the tests demonstrated several interesting effects, which caused much speaker inconsistency and hence poor recognition. Methods of encouraging speaker consistency are described. Subject comments on the tests were noted via a questionnaire. Subjects experienced various psychological reactions, from excitement to boredom and fatigue. Highest performance rates corresponded with highest concentration and speaking rhythm. These and other human factors observations are discussed in detail.

BACKGROUND

This particular research project into experimental design for speech recogniser performance assessment was devised from experiences gained during the completion of the European Economic Community's ESPRIT Project 449 - Investigation into the effective use of speech at the human-machine interface (1). In this project one of the consortium members carried out a long series of voice recogniser tests in a noisy industrial environment, with the aim of determining the performance capabilities of the technology in such harsh conditions. The results however were hugely disappointing, since poor experimental design and a lack of awareness of external factors influencing recogniser accuracy led to statistics which proved nothing about the performance in high ambient noise situations.

Briefly, the experimental design was as follows. Two speakers trained three separate vocabularies in a quiet laboratory environment, then tested each in a high noise workshop environment, in a range of background noise levels from 30dB to 120dB. Any substitution or deletion errors were noted. Analysis of the data proved difficult for several reasons. Firstly, tests were not performed at many intermediate noise levels but concentrated at the extremes of noise range. As for the noise levels that were tested, at some only a single repetition of the vocabulary was spoken whilst at others there was a number of repetitions. At all levels, there was an insufficient number of tokens spoken to achieve statistical significance. To further complicate matters, noise levels were measured only at the beginning and end of a test, although levels varied significantly throughout by as much as 60dB.

The major conclusion identified by the experiments was that sound experimental design was essential to achieve significant results. Using these
experiences as a base, a new set of experiments was planned to investigate a number of speaker related factors that affect recognition, with a carefully designed experimental methodology.

FACTORS INFLUENCING RECOGNITION ACCURACY

The difficulty in measuring speech recogniser performance is that there are very many variables that may affect performance to some degree. Over eighty factors have been identified which may affect recognition accuracy (2). Several of these refer to the implementation of the recognition algorithm and other aspects of hardware design which cannot be changed once a recogniser has been purchased. The rest refer to the conditions in which the recogniser is used, the environment in which the recogniser is situated, the vocabulary chosen and the human factors which affect performance.

The environment surrounding a speaker can affect the accuracy of voice recognition in two ways. Firstly, environmental noise and to a lesser extent, environmental factors such as room acoustics, humidity, temperature and pressure can affect the speech sound wave produced by the subject and received by the microphone. Secondly, the environment has an influence on the way the subject speaks, for example, in high ambient noise conditions, a subject will tend to speak louder than normal to maintain the usual speech to noise differential. Humidities or temperatures higher than normal may increase a subject's fatigue or cause a drying of the throat (3).

One of the most important decisions in experimental design concerns the choice of vocabulary, since this decision will have the greatest influence on whether subjects achieve high or lower rates of recognition. With an 'easy' vocabulary, the difference between best and worst performance may be very small, making analysis of the effect of dependent variables difficult to spot. With a more difficult vocabulary, the same dependent variables can become very clear. In the past, experiments have been performed with a wide range of vocabularies, from the ten digits up to hundreds phonetically similar words. Clearly, the larger the likelihood of misrecognition errors is higher for large vocabularies of similar sounding words than for smaller vocabularies of phonetically distinct words.

Vocabulary words also affect the consistency of the speaker. Certain words (such as 'either!') encourage inconsistency because they have several valid pronunciations, and subjects can forget which pronunciation was used during enrolment.

PROCEDURE

A set of experiments were conducted to investigate the effects of several speaker related factors on recognition performance. These factors were the age and sex of speakers, their accents and dialects and their experience with voice recognition technology.

A pool of male and female subjects were recruited from the company. The subjects were of all ages from late teens up to early sixties with a broad range of British accents. Experienced voice technology users and novices were equally matched.

Each subject was enrolled on a VOTAN VPC 2000 recogniser, and then asked to repeat a 64 word vocabulary 10 times, with the words presented in random order. Input gain was calibrated with each individual voice before enrolment. Acceptance level was default. Subjects were prompted with the words via a monitor. All substitution, insertion, deletion and rejection errors were recorded automatically. Background noise level was monitored throughout.

The vocabulary chosen was a practical vocabulary designed to operate the MS-DOS operating system, and contained commands along with the digits. Several statistical analyses were completed on the results. Recognition and error rates were calculated along with mean winning scores and standard deviations. Correlation coefficients were used to determine the effects of the dependent variables. Results showed that experience of using recognisers did significantly improve performance, whilst neither age nor accent were significant. It is interesting to report that the female speakers in the test achieved better performance than the males. Generally, in previous experiments the reverse has been true (4).
OBSERVATIONS

It is important for an observer to be present throughout all voice recognition tests. The primary aim of the observer is to monitor external factors that cannot be controlled, such as background noise, and to ensure that the tokens uttered correspond with the words the speaker was prompted for. The observer should also monitor the behaviour of the subject. The psychological state of a subject can influence performance in several ways. Stress or nerves can alter the voice of a subject. If a test commences with some form of enrolment, a nervous subject will record templates with a voice that is quite different from the voice used later in the test after the subject has relaxed. This relaxation after the initial 'excitement' of a new experience is particularly noticeable in novice users.

Relaxation can itself create problems, if subjects relax too much. If a test is too long, and a subject is merely repeating screen prompts, concentration soon begins to wane. (Indeed, in one test, on a particularly humid day in summer, one susceptible subject actually fell asleep.) In all cases in the experiment, the best performance rates were achieved by the subjects that were able to find and maintain a steady utterance rhythm. Subjects unable to find such a rhythm, suffered many more errors. These subjects invariably found the test unsettling, and would not sit still in their chairs. When interviewed after their tests, the reasons for the restlessness were generally a lack of interest in the task, impatience or outright boredom.

Some users, especially the novices, speak to the machine with a voice that is quite different from their normal speaking voice. They speak very properly, more loudly and over articulate, abandoning any natural accent for an imitation of BBC announcer English. This phenomenon can be likened to the telephone voices that some people with strong regional accents put on when they talk over a telephone line. Providing enrolment and practice are consistent, the effects of factors, whilst an easy vocabulary may not be affected at all.

GUIDELINES FOR EXPERIMENTAL DESIGN

Summarising the experiences gained during the experiments, the following tips are presented for future experiment designers. Firstly, ensure that you know exactly what it is to be investigated. Identify all unwanted factors that may influence performance. Keep as many as these under control as possible, monitor the factors beyond your control, if they vary significantly be prepared to abandon the test.

Identify the dependent variables that are to be investigated. Ensure that a full range are tested, with enough observations at each level to achieve statistical significance. Choose the test vocabulary carefully. Avoid words likely to cause speakers to be inconsistent. A difficult vocabulary will amplify the effects of factors, whilst an easy vocabulary may not be affected at all.

Avoid making the test too long for subjects, since this may cause a deterioration in motivation and an increase in fatigue. Be an observer in the experiments. Monitor subject behaviour and interview subjects after the test to discover their reactions. Document the details of subjects.

Analyse the data in every way that makes sense. This could include confusion matrices, graphs, recognition and error rates, mean winning scores, standard deviations, correlation analysis, hypothesis testing or statistical modelling.

REFERENCES

(1) 'ESPRIT Project 449: Investigation into the Effective Use of Speech at the Human-Machine Interface', Public Domain Report, 1986

(2) Wayne A Lea, 'What Causes Speech Recognizers to Make Mistakes?', IEEE ICASSP, 1982
