MICRO-MEASURES OF SPEECH RECOGNISER EFFECTIVENESS

P Kenne and M O'Kane
Faculty of Information Sciences and Engineering
University of Canberra

ABSTRACT - Speaker-independent speech recogniser performance measures are generally reported for performance averaged over all utterances by all speakers in a suitably large test database. In this paper we demonstrate that recognisers can perform quite differently on different speakers in a given test database. We examine the amount of variation that occurs across speakers for two different statistically-based recognisers - a Hidden Markov Model recogniser and a SPRITE recogniser.

Another issue that is hidden by averaged recogniser performance scores is the variation that can occur as a function of utterance length. Some recognisers perform better on shorter utterances for a given training database.

Finally, recogniser performance measures are examined. Some measures do not allow one to infer certain types of errors easily. Overguessing is one such example. Another measure is proposed and its performance is examined for effectiveness in highlighting different types of recogniser error.

INTRODUCTION

The work reported here is part of a larger study addressing the issue of characterising expected recogniser performance on test data for a given training database. The aim of this study is eventually to provide guidelines for the minimal database needed to achieve a pre-specified level of recognition for a given application.

In this paper we are concerned with an investigation of whether or not a simple averaged recogniser performance figure provides sufficient information about a recogniser's performance in new applications. We consider, for example, how the recogniser performs on different speakers. If the performance differs markedly for different speakers, is there some predictive explanation for this? Is the recogniser performance dependent on the length of the utterance? Can a analysis of recogniser performance on various measures give any guidance on the minimum training database needed to achieve a certain level of recognition?

In order to highlight these issues we examine the comparative performance of two types of recognisers on various measures.

MEASURES OF RECOGNISER PERFORMANCE

In the period 1982-1992 there has been a great deal of work done, particularly by the National Institute of Standards and Technology (NIST) and in projects funded by the American Defense Advanced Research Projects Agency (DARPA), aimed at creating massive databases for providing statistical material for training speech recognisers and separate databases for testing these recognisers. There has also been work done on developing measures to indicate performance of recognisers working over these standard test and training databases (Pallett, 1991). The main purpose these measures have been put to has been to summarise the performance of Hidden

Markov Model speech recognisers with various sophistications added to the basic model (Lee, 1989; Ramesh, Wilpon, McGee, Roe, Lee and Rabiner, 1991). The main concern is obtaining some reflection of average performance.

Commonly-used measures (cf Lee, 1989) are the following:

$$\text{word average} = \frac{\text{correct}}{\text{correct length}}$$

(This is often expressed as a percentage and called "Percent Correct")
and

$$\text{error rate} = 100 \frac{(\text{subs+dels+ins})}{\text{correct length}}$$

These measures were used in early versions of the work reported here but as there is the possibility of ambiguity in the definition of a substitution and as we were interested in worst-case phenomena we chose to retain Word Average but instead of Error Rate we calculate

$$\text{score average} = \frac{\text{correct}}{(\text{correct} + \text{ins} + \text{del})}$$

where a substitution becomes one insertion plus one deletion. As we wanted to examine length effects we added another measure

$$\text{length score} = 1 - \frac{\| \text{correct length} - \text{length of recognised string} \|}{\text{correct length}}$$

The score average rewards correct recognition and penalises misrecognition. This score penalises both types of misrecognition, insertion and deletion, and is thus harsher than scores that use substitution measures. The word average only provides a measure for the number correct and does not penalise mistakes of any type. The length score provides a measure of the length match.

RECOGNISERS COMPARED

The Hidden Markov Model (HMM) recogniser was developed in its current form by Lee (1989). This statistically-based recogniser has become popular worldwide and is the one most commonly referred to in the speech literature. The implementation used in this work was the Cambridge University HTK kit. There are however several other architectures for speech recognisers, for example, neural network recognisers and various recogniser architectures based on knowledge-based systems. We have developed a statistically-based recogniser known as SPRITE (O'Kane, Kenne, Landy and Atkins, 1991). The first experiment in this paper describes an investigation of the comparative performances of an HMM recogniser and a SPRITE recogniser, both trained using full-word models on the 55 male speakers from the TI digits training database (Leonard and Doddington, 1984). There are 77 utterances per speaker varying in utterance length from one digit to seven digits. There are, however, practically no strings of length 6.

It should be noted that the versions of the recognisers used in this experiment were minimalist in the sense that no grammars were formally incorporated into the recognisers. Also, in keeping with the worst-case approach, the category "silence" was removed before the recogniser scores were

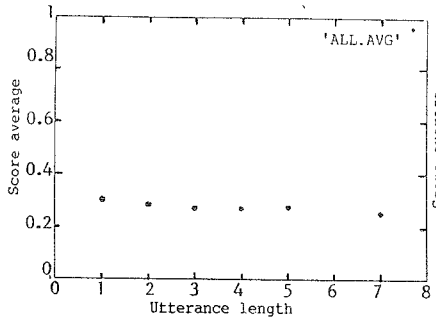

Figure 1a: Score average as a function of utterance length for HMM recogniser

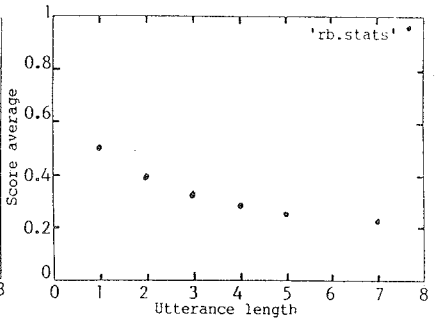

Figure 1b: Score average as a function of utterance length for SPRITE recogniser

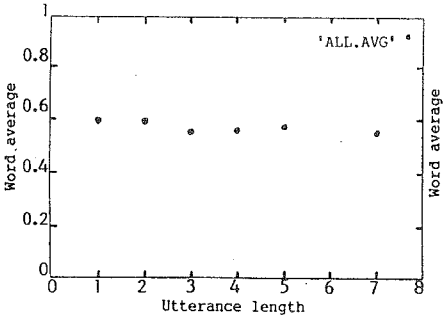

Figure 2a: Word average as a function of utterance length for HMM recogniser

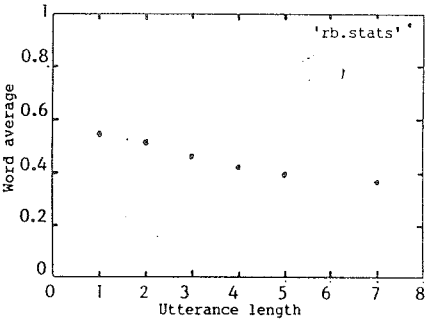

Figure 2b: Word averagte as a function of utterance length for SPRITE recogniser

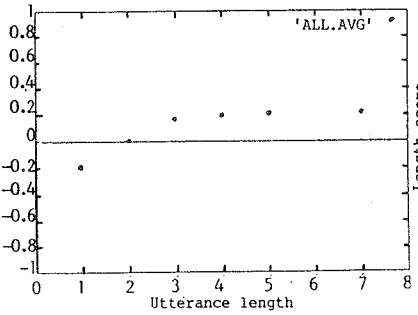

Figure 3a: Length score as a function of utterance length for HMM recogniser

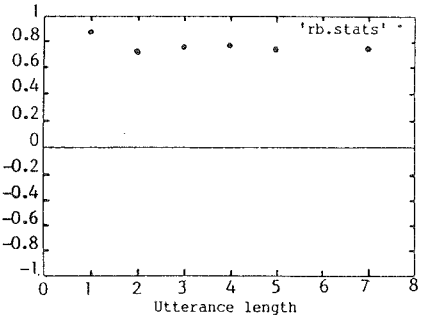

Figure 3b: Length score as a function of utterance length for SPRITE recogniser

calculated although both types of recognisers reported "silence" as an allowable primary result category.

We tested both recognisers on 23 speakers from the TI male test database. There are again 77 utterances for each of these speakers. Figures 1a-b, 2a-b and 3a-b give the results averaged across all speakers for score average, word average and length score plotted as a function of correct length. Table 1 gives the performance for each recogniser for each measure averaged across all test utterances for all test speakers.

	Score Average	Word Average	Length Score
HMM	2.89	5.85	0.08
SPRITE	3.51	4.64	0.80

Table 1: Comparative results for HMM and SPRITE recognisers averaged across the 23 test speakers (77 utterances/speaker).

Considering the averaged results, the HMM recogniser scores significantly better than the SPRITE recogniser on the word average measure but slightly worse than the SPRITE recogniser on the score average and significantly worse on the length score. What does this mean? The best way to examine this is to go back to the definition of scores and the comment that was made above. The word average reflects the number correct regardless of the mistakes made. The fact that the HMM tends to have a high word average and low length average means that it tends to propose strings with a lot of possibilities in them at the expense of being over-long. With the SPRITE recogniser on the other hand, the length average is close to 1. In other words, it attempts to recognise a string of approximately the right length at the price of actually making considerable substitution errors.

Let us turn to the question of how the recognisers perform as the length of the string that they are attempting to recognise varies. The score average and word average are length-independent for the HMM recogniser while they are both length-dependent (shorter scores better) for the SPRITE recogniser. The length score is relatively length-independent for the SPRITE recogniser and length-dependent for the HMM recogniser (shorter scores worse).

COMPARISON FOR DIFFERENT SPEAKERS

We examined the performance of both recognisers for each of the 23 speakers from the test database. Both recognisers perform best (by all three measures), by some margin, on the speaker "bc". The graphs of score average versus utterance length for "bc" are given in Figures 4a-b. Good performance on speaker "bc" is not too surprising as the majority of speakers in the TI digits training database are from southern states of the USA and the average age of the speakers in the training set is 30. Speaker 'bc" is from Texas and is aged 23. On the other hand both recognisers do badly on the speaker "ga" (worst performance for SPRITE recogniser, fourth-worst performance for HMM recogniser). Speaker "ga" is 70 and is from New York City. A more detailed description of recogniser performance for individual speakers can be found in a paper by O'Kane and Kenne (1992).

The other issue investigated was the variation of word average versus score average (see Figure 5a-b) for all 23 speakers in the test set. The results were averaged across all utterances of the same length for each speaker and for both versions of the recogniser. The different symbols on the diagram refer to utterance length. There is an overall linear relationship between word average and score average for the HMM recogniser. For the SPRITE recogniser there appears to be a linear relationship between word average and score average and for each utterance length. As the utterance length decreases, the line for that length goes increasingly close to going through the point (1,1) on the graph. It is particularly noteworthy that neither the HMM line nor any of the SPRITE lines go through the point (1,1), the point of no errors. This implies that for both recognisers larger amounts of training data are needed to achieve perfect recognition. For the SPRITE recogniser, more training data are needed to achieve perfect recognition on long utterances than on short utterances

CONCLUSION

An examination of the micro performance of recognisers is useful in highlighting weaknesses in the recognisers and thereby giving indications of possible improvements. It can also highlight comparative strengths and weaknesses of different types of recognisers that might lead to the recognisers being used in a complementary manner. Furthermore it provides indications of lower bounds for amounts of recogniser training data needed for perfect recognition.

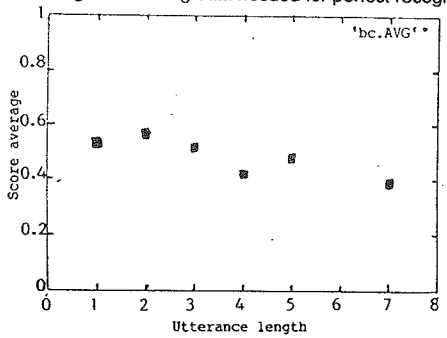

Figure 4a: Score average as a function of utterance length for HMM recogniser for speaker "bc", the "best" speaker in the test set

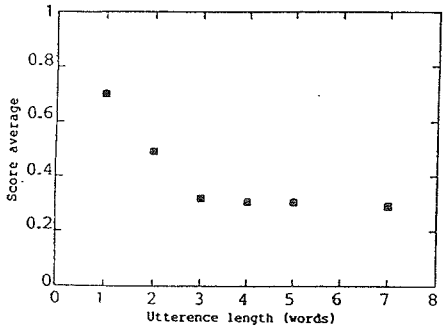

Figure 4b: Score average as a function of utterance length for SPRITE recogniser for speaker "bc" the "best" speaker in the test set.

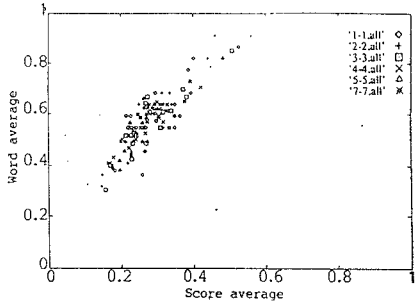

Figure 5a: Word average versus score average for HMM recogniser averaged for all speakers from test set. Results for each speaker for each utterance length are averaged to obtain points on the graph

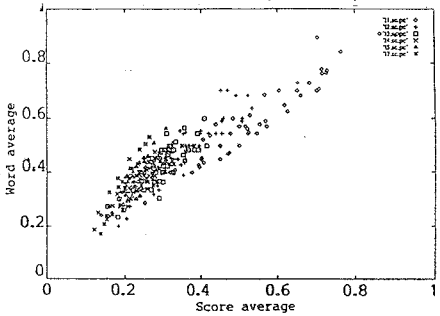

Figure 5b: Word average versus score average for SPRITE recogniser averaged for all speakers from test set. Results for each speaker for each utterance length are averaged to obtain points on the graph

REFERENCES

Lee, K-F, (1989) *Automatic Speech Recognition - The Development of the SPHINX System*,Kluwer, Norwell.

Leonard, R G and Doddington, G, (1984) *A database for speaker-independent digit recognition,* Proceedings ECASSP'84, Vol 3, 42.11

O'Kane, M and Kenne, P, (1992) *A Review of Speech Modelling Research and Applications - Measuring and Comparing Recognisers*, Proceedings Communications'92, Sydney.

O'Kane, M , Kenne, P, Landy D and Atkins, S, (1991) *Generalising from Single-Speaker Recognition in a Feature-Based Recogniser,* Proceedings of Eurospeech'91, Genova, 409-412.

Pallett, D (Chair)(1991) *DARPA Resoruce Management and ATIS Benchmark Test Poster Session,* Proceedings Speech and Natural Language Workshop, Pacific Grove, Morgan Kaufmann, 49-137.

Ramesh, R, Wilpon, J G, McGee, MA, Roe, D B, Lee C H and Rabiner, L R, (1991) *Speaker Independent Recognition of Spontaneuously Spoken Connected Digits,* Proceedings of Eurospeech'91, Genova, 17-20.