# PERFORMANCE OF HUMANS ON AN ISOLATED WORD SPEECH RECOGNITION TASK

S. J. Cox (1), P. W. Linford (1), K. O. Chichlowski (1) and R. D. Johnston (2)

(1) School of Information Systems, University of East Anglia, Norwich NR4 7TJ.
(2) Speech Applications Division, British Telecom Laboratories, Ipswich IP5 7RE.

## INTRODUCTION

In this paper, we report on a pilot experiment in which human listeners were given a speech recognition task which consisted of identifying isolated utterances of the alphabet. The database used was the BT CONNEX database which has been made generally available for research purposes in UK universities and industry.

We had several motivations for this experiment. Firstly, considerable effort within the speech recognition community is devoted to improving the performance of speech recognition systems and in some cases, very high levels of performance (i.e. very low error-rates) have been reported. In cases where a significant effort has been expended on optimising a system, it would be useful to know how much room there is for further improvement, since the 'law of diminishing returns' means that such an improvement is likely to be expensive to obtain. It is therefore of considerable interest to obtain an upper bound for the performance on a particular recognition task, and we assume that this upper bound can be estimated by measuring the performance of a human with normal hearing who is a native speaker of the language he is hearing.

Our wider objective is to lay the foundations for developing methods to calibrate the difficulty of different speech recognition tasks using a reference distortion approach which has been developed for characterising telecommunication channels. Such methods avoid the many problems associated with having to characterise speech used as the 'raw material' for any test. Instead, they depend upon being able to devise a means for controllably and representatively impairing speech quality to provide a calibrated reference system. Using such a system it is then possible to rate the performance of different speech technology systems over a wide range of speech and speaker types.

A further objective was to make a list of utterances in the CONNEX database which were corrupt in some way and whose inclusion in speech recognition experiments was open to question. Such a list was made after the database was first recorded, but it was felt to be useful to have an independent check.

Finally, it should be stressed that this was a pilot experiment to give us an initial experience in designing, conducting and analysing the results of such a test and to enable us to establish the infrastructure necessary to run tests. For this reason, the test was not as large or as comprehensive as we would have ideally liked.

## 2. THE DATABASE

The BT CONNEX database consists of utterances from 104 speakers each speaking 3 utterances of the 26 letters of the (British) alphabet. It was recorded under the following conditions:

- recording environment was a soundproof booth
- high-quality headset microphone used
- recording bandwidth 100 Hz–8 kHz
- 16-bit A/D converter
- sampling-rate 20 kHz.

The recording was done under computer control by a system which prompted the speaker to say a letter of the alphabet by displaying it on a VDU, and which then began recording to disk for a fixed period of 2 seconds. The order in which the 78 letters were presented to the speaker was randomised. During a preliminary session, the gain of the recording system was adjusted to accommodate the average level of each speaker's voice, but there was no subsequent energy or amplitude normalisation of the recorded speech. Each utterance was recorded to a separate file.

## 3. EXPERIMENTAL DESIGN

### 3.1 Signal Conditions

It was decided to run the test at two different bandwidths: bandwidth 1 (BW1), which was the original full bandwidth of 100 Hz-8 kHz and bandwidth 2 (BW2), which was telephony bandwidth i.e. 300 Hz-3.4 kHz. This would lay down benchmark results for noise-free conditions at wide and restricted bandwidth and also enable comparisons with results from automatic recognition algorithms done at these bandwidths. A more comprehensive test would include results done at different listening levels and signal-to-noise ratios.

### 3.2 Allocation of utterances to listeners

The following considerations governed the experimental design:

- The number of listeners that it was practicable to test
- The nunber of sessions that it was practicable to ask each listener to attend
- The number of words a listener could hear in a session without danger of fatigue
- A 'balanced design' was used. The use of a balanced design meant that various factors (speakers, listeners, presentation sequence and utterance type) could be tested against each other in the subsequent analysis.

The main features of the chosen design were as follows:

- 26 listeners
- Each listener heard 390 utterances at BW1 and the same 390 at BW2
- Order of presentation of classes randomised
- Utterances played to a listener in 3 sessions of approximately 20 minutes each session
- Listener's sessions separated by at least one day
- Each listener heard data from each speaker
- Each listener heard a unique set of utterances, apart from a small number of utterances heard by several listeners

The selection of utterances to be played to listeners was made as follows. For his first session, Listener 1 was allocated a set of 104 utterances consisting of 4 repetitions of the alphabet. Each utterance came from a different speaker and was the first utterance of the three each speaker had provided. Listener 2 was then allocated 4 repetitions of the alphabet in which the speakers spoke different letters from those allocated to Listener—again, these were "first" utterances. This process was continued until all 26 listeners had been allocated utterances. In addition, each listener was allocated another 26 utterances, each one taken from the list of another listener. The resulting 130 utterances were heard by listeners at both bandwidths, making a total of 260 utterances in a session. The second and third sessions were constructed in exactly the same way except that the second and third utterances from the speakers were used. No utterance was played out more than twice and only about 30% were played twice.

### 3.3 Selection of listeners

Because this was essentially a pilot experiment, the criteria for selecting listeners were confined to the following simple requirements:

- listeners should have no history of hearing problems (however, we did not test listeners' hearing ourselves)
- listeners should be native English speakers
- there should be equal numbers of males and females
- listeners should be in the age-range 18–30

## 4. EXPERIMENTAL PROCEDURE

A block diagram of the arrangement used to play out utterances and record listener's responses is shown in Figure 1. The test was controlled by the PC which both recorded listeners responses and controlled the outputting of utterances from the filestore. The program displayed a screen with 28 "buttons", one for each letter of the alphabet plus an asterisk (to be used for unrecognisable utterances or "rejections") and a "START" button (to begin a session). A "button" could be pressed by dragging a pointer (controlled by a mouse) onto the button and clicking, in the usual way. The "buttons" were displayed in five rows and six columns on the screen with the START buttom in the top LH corner and the asterisk in the bottom RH corner. In between, the buttons were displayed in alphabetical order. The advantage of this arrangement is that subjects can easily find the button they want to choose. The disadvantage is that some commonly confused letters are then adjacent on the screen e.g. "B", "C", "D" and "E", "J" and "K". This means that it is difficult to distinguish genuine classification errors from cases in which the wrong button has been inadvertently selected (see Section 5.2).

The selection of a button recorded the appropriate response in a file on the PC and also signalled the computer to play out the next utterance in the list, so that the listener was allowed to hear the utterance only once but was given as much time as he wanted to reach a decision about its classification.

The gain of the final amplifier was fixed throughout the experiments and was determined by an informal test in which listeners (not subjects in the test) set the gain to a level which was comfortable for them. The gain used was an average of their settings, which were very similar. Prior to the first session, listeners were given a practice run (in which they heard 10 utterances at each bandwidth) and were given the option of a practice run before each session. They were advised to select the asterisk only if the utterance was completely unrecognisable.

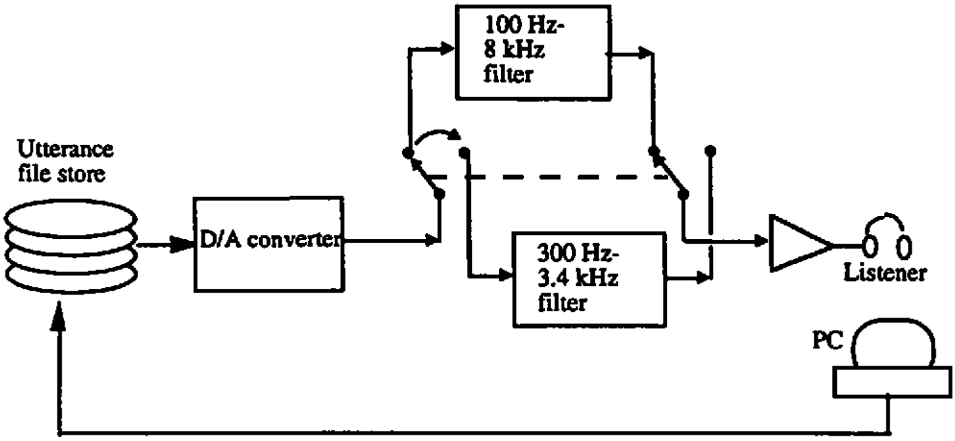

**Fig 1: Experimental arrangement**

## 5. ANALYSIS OF RESULTS

### 5.1 Average error-rates over all listeners

Because every listener heard a different subset of the database (with a very small overlap), it is not strictly correct to average their accuracy figures and claim that it is an estimate of the overall accuracy. However, each listener heard a balanced set of examples from each speaker and so if we can assume that "good" and "bad" *utterances* are highly correlated with *speakers*, interactions between good/bad utterances and good/bad listeners will average out, and the error averaging procedure is not unreasonable. The assertion that these interactions will average out needs to be established by a full analysis of variance on the data which we have not yet completed.

The error figures (averaged over listeners) for the full CONNEX dataset are presented below:

| | Error-rate 1 (%)<br>(subs only) | Error-rate 2 (%)<br>(subs+rejs) |
|---|---|---|
| High bandwidth | 2.27 | 2.96 |
| Low bandwidth | 2.52 | 3.24 |

**Table 1: Error-rates averaged over all listeners for full CONNEX dataset**

In Table 1, error-rate 1 is the error-rate when "rejections" (utterances which the listener was unable to identify) are discounted and error-rate 2 is that obtained when rejections are counted as "substitutions" (utterances incorrectly identified). The total number of utterances played out to listeners was 10 140 of which 8112 (the total number of utterances in the database) were different.

### 5.2 Identification of "faulty" utterances

The errors reported in section 5.1 constitute a total of 271 mis-recognised utterance files (both substitutions and rejections). After all listener sessions were complete, each of these mis-recognised files was replayed several times over headphones by one researcher in an attempt to establish what had lead to the mis-recognition. In the few cases where it was not clear what the problem was, the file was heard by several researchers who arrived at a consensus view.

It was stated in section 1 that one of the objectives of this experiment was to identify "faulty" utterance files in the CONNEX database. Four different faults were noted within the set of mis-recognised files:

1. nothing recorded in file (54 utterances)
2. speaker uttered a different word from the one prompted for (26 utterances)
3. utterance truncated because speaker began speaking too soon, or by recording timeout, or by faulty endpointing (13 utterances)
4. utterance so loud or quiet as to make classification difficult (22 utterances)

In practice, fault 1 is the only fault in the above list which can be unambiguously identified. Fault 2 was identified only when it was very obvious e.g. when the speaker had mis-read "L" as "One" or "V" as "Why" from the prompt screen. Note that pronouncing "Z" as "Zee" was regarded as a fault. Faults 3 and 4 were present in varying degrees but it was generally clear when one or both of these faults (rather than any phonetic irregularity in the utterance) had lead to a mis-recognition.

In addition, there were 32 responses (not utterances) in which it was suspected that there had been a listener error i.e. the listener had selected the wrong screen "button". The reason for this suspicion was that in each of these cases, the utterance was clear but the listener's response corresponded to selecting a button which was one position removed from the correct button. However, it was noticed that utterances giving rise to possible "button errors" when heard at one bandwidth tended to also give the same error at the other bandwidth, which cast serious doubts on the theory that these responses were button errors. It is difficult to explain the errors on these utterances; they were unambiguous to the researcher who heard them after the tests, which suggests a button error, but it seems unlikely that a listener would make the same motor error twice. To keep the error-estimate conservative, these responses have been counted as genuine mis-recognitions.

### 5.3 Results when "faulty" utterances are excluded

When the 115 "faulty" utterances are excluded from the results, the error-rates are as shown in Table 2:

| | Error-rate 1 (%)<br>(subs only) | Error-rate 2 (%)<br>(subs+rejs) |
|---|---|---|
| High bandwidth | 1.70 | 1.89 |
| Low bandwidth | 1.98 | 2.15 |

**Table 2: Error-rates averaged over all listeners ("bad" utterances excluded)**

After exclusion of "bad" utterances, the total number of utterances considered in the analysis was 10 012 of which 7997 were different.

The confusion matrix for the low bandwidth case is given as Table 3. Rows represent the class of the stimulus and columns the class of the response (note that the response class '*' represents rejection). The high bandwidth confusion matrix is not included here for reasons of space, but is very similar. Many confusions shown here are predictable, such as 'E' set members, M/N and S/F. However, three frequently occuring confusions would not be predicted from a homogeneous group of RP speakers: A/E, A/I, G/J. It is probable that these confusions reflect the large number of speakers from Scotland and Northern Ireland in the dataset.

### 5.4 Comparison of results at high and low bandwidths

Listeners heard the same utterances at the two bandwidths (randomised over the sessions) and so a comparison of error-rates at the two bandwidths is valid. (Strictly, one should compare this variation with the variation in response from session to session when a listener is given the same set of stimulii in each session, but we assume that this variation would be small.)

The *sign test* is an appropriate test to apply in this case. The null hypothesis $H_0$ is that the error-rate of a listener is the same on both bandwidths. Let the i'th listener make $H_i$ errors on the high bandwidth data and $L_i$ errors on the low bandwidth data. Let $D_i = \mathrm{sgn}(H_i - L_i)$ and discard all zero values of $D_i$. Under $H_0$, $P_i$, the number of positive values of $D_i$, has a Binomial distribution with parameters N (the number of non-zero values of $D_i$) and underlying probability p=0.5.

Subtracting the high bandwidth total errors column from the low bandwidth total errors column in Table 3.4 gives 5 zeroes, 6 positive and 15 negative numbers. The likelihood of observing this distribution under $H_0$ is 0.0392. Since we are testing for a *difference* in the values on high and low bandwidths, we use a two-sided test and state that the probability of observing this difference by chance is 0.0784 or about 8%. Hence there is some evidence of an effect due to reduced bandwidth but further testing would be required to be confident of this assertion. What is perhaps of greater significance is that practically speaking, there was very little difference in the error-rates at the two bandwidths.

## 6 DISCUSSION AND FUTURE WORK

The main results to emerge from this pilot study are:

(1) On the task of recognition of the alphabet using data recorded under "ideal" conditions, human performance was high, with a conservatively estimated error-rate of under 2%. Some error-rates reported in the literature for automatic speech recognition systems classifying the "test-set" of this data (utterances from 52 of the speakers, 26 male and 26 female) are:

14.8% [1] using a hidden Markov model with 3-component mixture densities

12.4% [2] using a multi-layer perceptron with 50 hidden units—this is the best result from a set of 35 results using different classifiers

The result of 2.8% reported in [3] is for a "multi-speaker" experiment in which utterances from all 104 speakers were included in the training- and test-sets and is therefore not

|  | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | • |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A : | 364 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B : | 2 | 385 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 |
| C : | 0 | 0 | 388 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| D : | 0 | 1 | 0 | 384 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E : | 24 | 2 | 0 | 1 | 348 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 7 |
| F : | 0 | 0 | 0 | 0 | 1 | 384 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G : | 0 | 0 | 0 | 0 | 0 | 0 | 369 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| H : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 389 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I : | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 374 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| J : | 0 | 0 | 0 | 1 | 0 | 0 | 11 | 0 | 0 | 377 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 382 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| L : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 350 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 370 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| N : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 383 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| O : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 386 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P : | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 386 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Q : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 377 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 388 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S : | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 381 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T : | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 384 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| U : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 387 | 0 | 0 | 0 | 0 | 0 | 0 |
| V : | 1 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 368 | 0 | 0 | 0 | 0 | 2 |
| W: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 384 | 0 | 0 | 0 | 1 |
| X : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 385 | 0 | 0 | 0 |
| Y : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 389 | 0 | 0 |
| Z : | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 381 | 1 |
|  | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | • |

Table 3: High bandwidth confusion matrix

comparable with tests in which the listeners had never heard the speakers before. The results quoted above were not for systems that attempted to *optimise* performance on this data and are not "state of the art" recognisers. However, even if recent advances in speech-recognition techniques were utilised and the recognisers optimised for this data, it seems unlikely that the very large increase in performance needed to approach human performance would be obtained.

(2) When the bandwidth was restricted to telephone-bandwidth, the performance dropped only very slightly.

(3) Mis-recognitions could be broadly divided into those that could be predicted from an RP accent model and those which it is assumed were due to the large number of Irish and Scottish accents in the database.

In addition, we have produced a more definitive list of "faulty" utterances from the CONNEX database based on human misrecognitions—this list is available from S.J.Cox. (sjc@sys.uea.ac.uk).

The next step is to find a method for reliably impairing the speech output in such a way that the "recognisability" of speech data is smoothly and monotonically reduced as the level of impairment increases. Such an impairment unit would then form the basis for calibrating databases and testing speech recognisers.

## ACKNOWLEDGMENT

The work reported here was undertaken at UEA with the support of a contract from BT Laboratories.

## REFERENCES

[1] S.J.COX & J.S.BRIDLE, Simultaneous Speaker Normalisation and Utterance Labelling using Bayesian/Neural Net techniques. *Proc IEEE Conf. on Acoustics, Speech and Signal Processing, Albuquerque, 1990*

[2] P.W.LINFORD & G.D. TATTERSALL, Non-linear Time Normalization of Utterances for Speech Recognition using MLP's. *Proc Inst. of Acoustics 1990 Autumn Conference, Vol 12: Part 10, pp 291-297.*

[3] P.C.WOODLAND, Hidden Markov Models using Vector Linear Prediction and Discriminative Output Distributions. *Proc IEEE Conf. on Acoustics, Speech and Signal Processing,, San Francisco, 1992*