S10b.5

SOME STATISTICAL ISSUES IN THE COMPARISON OF SPEECH RECOGNITION ALGORITHMS

L. Gillick[†]

Dragon Systems, Inc.,
Chapel Bridge Park,
90 Bridge Street,
Newton, MA 02158, USA

S.J. Cox

RT5233,
British Telecom Research Laboratories,
Ipswich IP5 7RE, UK.

ABSTRACT

In the development of speech recognition algorithms, it is important to know whether any apparent difference in performance of algorithms is statistically significant, yet this issue is almost always overlooked. We present two simple tests for deciding whether the difference in error-rates between two algorithms tested on the same data set is statistically significant. The first (McNemar's test) requires the errors made by an algorithm to be independent events and is most appropriate for isolated word algorithms. The second (a matched-pairs test) can be used even when errors are not independent events and is more appropriate for connected speech.

1 INTRODUCTION

The speech recognition literature currently abounds with descriptions of novel or improved algorithms for speech recognition. It is common practice for researchers to test two or more algorithms together and then to make claims for their relative efficacy on the basis of the test results. However, these claims are seldom backed by evidence that any difference in performance is statistically significant; indeed, most papers show an almost complete lack of awareness of the importance of comparing results of experiments in a way that takes account of variability and uncertainty in a principled manner. In this paper, we present some statistical ideas and techniques that will make it possible to perform such comparisons on algorithms (or systems) that recognise isolated words and connected or continuous speech. We hope to thereby encourage researchers who are reporting empirical results to use statistical measures in summarizing their findings and drawing conclusions.

We concentrate on methods in which the algorithms are tested on the same data set. Algorithms are often compared by testing them with the same data because by forcing the test items to be the same, the results then reflect differences between the algorithms rather than any accidental differences in the difficulty of the test items in independent data sets. However, the constraint of testing different algorithms on the same data set calls for a more sophisticated statistical approach than that required if each algorithm were tested on an independent set.

[†]Work supported in part by DARPA under contract number N00039-86-C-0307

1.1 Notation

We shall use capital letters throughout to denote random variables (RVs) and lowercase letters for scalars or observed values of random variables. An exception to the above rule is an estimate of a parameter which, although it is an RV, we denote by a cicumflexed lower-case letter.

2 A SIMPLE APPROACH

Suppose there are two algorithms, A_1 and A_2, which are presented with a sequence of labelled utterances $\{u_i\} = u_1, u_2, \ldots, u_n$ for recognition. By 'recognition', we mean that the algorithms make a decision about the label of each $\{u_i\}$ which is either correct or incorrect. We suppose that the $\{u_i\}$ are representative of some larger population of utterance sequences. We also assume that the $\{u_i\}$ are isolated utterances of e.g. syllables, words or phrases. Now suppose that the true (but unknown) error-rates of A_1 and A_2 are respectively p_1 and p_2. Our task is to decide whether there is enough evidence to conclude that either $p_1 > p_2$, $p_1 = p_2$ or $p_1 < p_2$.

Define the random variable X_i^j as follows:

$$\begin{aligned} X_i^j &= 0 \quad \text{when } A_j \text{ labels } u_i \text{ correctly} \\ &= 1 \quad \text{when } A_j \text{ labels } u_i \text{ incorrectly} \end{aligned}$$

It is reasonable to assume that $S^j = \sum_{i=1}^{n} X_i^j$ follows a binomial distribution $\mathcal{B}(n, p_j)$ as long as the errors are independent events. The maximum likelihood (ML) estimate of p_j is $\hat{p}_j$:

$$\hat{p}_j = \frac{S^j}{n} \tag{1}$$

The variance of $\hat{p}_j$ is σ_j^2:

$$\sigma_j^2 = \frac{p_j(1-p_j)}{n} \tag{2}$$

We would like to test the null hypothesis:

$$\mathbf{H_0}: \qquad p_1 = p_2 = p \tag{3}$$

which is equivalent to the hypothesis that $d = p_1 - p_2 = 0$. Under $\mathbf{H_0}$, the ML estimate of d is $\hat{d} = \hat{p}_1 - \hat{p}_2$, with associated variance σ_d^2:

$$\sigma_d^2 = Var(\hat{p}_1 - \hat{p}_2) \tag{4}$$

If we can assume that $\hat{p}_1$ and $\hat{p}_2$ are independent, equation 4 can be written as:

$$\sigma_d^2 = Var(\hat{p}_1) + Var(\hat{p}_2) = \sigma_1^2 + \sigma_2^2 \tag{5}$$

CH2673-2/89/0000-0532 $1.00 © 1989 IEEE

and σ_d^2 can be estimated by:

$$\hat{\sigma}_d^2 = \frac{2\hat{p}(1-\hat{p})}{n} \qquad \text{if } \mathbf{H_0} \text{ is true} \tag{6}$$

where the ML estimate of p is $\hat{p} = (\hat{p_1} + \hat{p_2})/2$. Then if n is large enough and $\mathbf{H_0}$ is true, the distribution of the statistic

$$W = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}} \tag{7}$$

tends to a Normal distribution with zero mean and unit variance, $\mathcal{N}(0,1)$. To test the null hypothesis, we compute $P = 2\Pr(Z \geq |w|)$ where Z is a random variable with distribution $\mathcal{N}(0,1)$ and w is the realized value of W. If it is found that $P < \alpha$ for a chosen significance level α, $\mathbf{H_0}$ is rejected (typical values of α are 0.05, 0.01 or 0.001). Notice the factor of 2 in the computation of P; W can lie on either side of the mean (zero) depending on the values of $\hat{p_1}$ and $\hat{p_2}$, and we include both possibilities (a *two-tailed* test).

2.1 An example

Let us take an example to illustrate the above procedure. In a recent test, A_1 and A_2 made 72 and 62 errors respectively on a test-set of 1400 utterances. Hence $n = 1400$, $\hat{p_1} = 0.0514$, $\hat{p_2} = 0.0443$ and substituting into equation 7 gives $w = 0.8853$. Using $z = |w| = 0.8853$ in standard tables of the unit Normal distribution, we obtain $P' = \int_{-\infty}^{z} \phi(t)dt = 0.812$, where $\phi(t)$ is the density corresponding to $\mathcal{N}(0,1)$. However, we require $P = 2\int_{z}^{\infty} \phi(t)dt = 2(1-P')$, giving a P-value $P = 0.376$. The conclusion is that $\mathbf{H_0}$ cannot be rejected and the difference in performance might well be due to chance effects.

2.2 Comments on the validity of the simple approach

The above analysis is correct if the assumption that $\hat{p_1}$ and $\hat{p_2}$ are independent is valid. Unfortunately, this cannot be true when A_1 and A_2 are tested on the same data-set because if the algorithms are similar, they may have many errors in common. We might approach this problem by attempting to estimate the correlation of $\hat{p_1}$ and $\hat{p_2}$. The difficulty with this idea is that we have only one data-set and any partioning of this data will increase the estimates of the variance of $\hat{p_1}$ and $\hat{p_2}$. A more direct and elegant solution is offered by McNemar's test.

3 McNEMAR'S TEST

The joint performance of the two algorithms can be summarised in a 2×2 table as follows:

		A_2 Correct	A_2 Incorrect
A_1	Correct	N_{00}	N_{01}
	Incorrect	N_{10}	N_{11}

where:

N_{00} = No of utterances which A_1 classifies correctly, A_2 classifies correctly
N_{01} = No of utterances which A_1 classifies correctly, A_2 classifies incorrectly
N_{10} = No of utterances which A_1 classifies incorrectly, A_2 classifies correctly
N_{11} = No of utterances which A_1 classifies incorrectly, A_2 classifies incorrectly

Note that $n = N_{00} + N_{01} + N_{10} + N_{11}$. By analogy with the N_{ij}'s, define q_{ij}'s:
$q_{00} = \Pr(A_1$ classifies u_i correctly, A_2 classifies u_i correctly) etc. Hence $E(N_{ij}) = nq_{ij}$. Now observe that $p_1 = q_{10} + q_{11}$ and $p_2 = q_{01} + q_{11}$. Therefore $\mathbf{H_0}$ is equivalent to $\mathbf{H_0^1}$: $q_{01} = q_{10}$. Defining $q = q_{10}/(q_{01} + q_{10})$, a further equivalent null hypothesis is $\mathbf{H_0^2}$: $q = \frac{1}{2}$.

The parameter q represents the conditional probability that A_1 will make an error on an utterance given that only one of the two algorithms makes an error. The null hypothesis $q = \frac{1}{2}$ represents the assertion that, given that only one of the algorithms makes an error, it is equally likely to be either one. It is now plausible that to test $\mathbf{H_0}$, it should only be necessary to examine the utterances on which only one of the algorithms made an error. No information about the *relative* performance of A_1 and A_2 is available from utterances on which they are both right or both wrong.

If we condition on the number of utterances $K = N_{10} + N_{01}$ on which only one algorithm made an error, then for the observed $K = k$, N_{10} has a $\mathcal{B}(k,q)$ distribution. Furthermore, under $\mathbf{H_0}$, N_{10} has a $\mathcal{B}(k,\frac{1}{2})$ distribution. The null hypothesis is thus tested by applying a two-tailed test (as in section 2) to the observation of a random variable M drawn from a $\mathcal{B}(k,\frac{1}{2})$ distribution:

$$\begin{aligned} P &= 2\Pr(n_{10} \leq M \leq k) && \text{when } n_{10} > k/2 && (8)\\ &= 2\Pr(0 \leq M \leq n_{10}) && \text{when } n_{10} < k/2 && (9)\\ &= 1.0 && \text{when } n_{10} = k/2 && (10) \end{aligned}$$

The probabilities can be computed directly as follows:

$$\begin{aligned} P &= 2\sum_{m=n_{10}}^{k} \binom{k}{m} \left(\frac{1}{2}\right)^k && \text{when } n_{10} > k/2 && (11)\\ &= 2\sum_{m=0}^{n_{10}} \binom{k}{m} \left(\frac{1}{2}\right)^k && \text{when } n_{10} < k/2 && (12) \end{aligned}$$

or alternatively, tables of the Binomial distribution may be used. As in section 2, $\mathbf{H_0}$ is rejected when P is less than some significance level α. If k is large enough ($k > 50$) and n_{10} is not too close to k or 0, a Normal approximation to the exact Binomial probability may be used. Under $\mathbf{H_0}$ and conditional on $K = k$, $E(N_{10}) = k/2$ and $Var(N_{10}) = k/4$. Then let

$$W = \frac{|N_{10} - \frac{k}{2}| - \frac{1}{2}}{\sqrt{\frac{k}{4}}} \tag{13}$$

which should be approximately $\mathcal{N}(0,1)$ under $\mathbf{H_0}$. Compute the P-value $P = 2\Pr(Z \geq w)$ (where Z is a random variable with distribution $\mathcal{N}(0,1)$ and w is the realized value of W), and reject $\mathbf{H_0}$ if $P < \alpha$, where α is the chosen significance level.

The $-\frac{1}{2}$ in the numerator of equation 13 is a continuity correction factor [1]. This latter form of the test is equivalent to the χ^2 test of McNemar [4].

3.1 Examples using the same test set

Let us now re-examine the data tested in section 2.1, where the errors were assumed to be independent. The distribution of the errors was:

		A_2	
		Correct	Incorrect
A_1	Correct	1325	3
	Incorrect	13	59

Hence $k = 16$, $n_{10} = 13$ and P is computed directly from equation 11 as 0.0213 (using the Normal approximation gives $P = 0.0244$). The observed difference would arise by chance on about 2 % of occasions (c.f. 37.6 % of occasions if it is assumed that $\hat{p}_1$ and $\hat{p}_2$ are independent), so there is evidence of a genuine difference.

In this case, the application of McNemar's Test increases our confidence that we are observing a genuine difference in the performance of two algorithms. It is instructive to compare different distributions of incorrectly classified utterances and their associated values of P (exact P-values are given, with P-value under Normal approximation in brackets):

	A_2	
A_1	1266	62
	72	0

$P = 0.437$ (0.437)

	A_2	
A_1	1328	0
	10	62

$P = 0.0020$ (0.0044)

3.2 Non-independent errors

McNemar's test is applicable when the errors made by an algorithm are independent, a condition which is certainly met in an isolated word algorithm which does not make use of any context. If the algorithm requires isolated word input but uses any sort of language model, the errors will no longer be independent. Clearly, useful information on the performance of the acoustic modelling is obtained by disabling the language-model and applying McNemar's test to the individual words. Some information on the combined performance of acoustic modelling and language model could be obtained by supplying the algorithm with enough 'left context' to ensure that the probability of error on a particular word is independent of any earlier errors. For example, before the performance on word k in a sentence is recorded, supply the correct labels of words $k-l, k-l+1, \ldots, k-1$ to the algorithm, where l is large enough for the effect of errors on any previous words to be neutralised. The errors are now independent and the effect of the language model is included. Given that users will probably want to do immediate error correction of each misrecognised word in systems of this sort, this is a reasonable strategy.

Current algorithms for recognising connected speech consider the whole spoken phrase before deciding upon the most likely 'explanation' (labelling) of the acoustic input. The errors are therefore highly inter-dependent and it would be wrong to attempt to compare performance on segments of a phrase which were in error. However, if each spoken phrase is reasonably short (i.e. a few words), it can be considered as an entity which is either recognised correctly or incorrectly, and McNemar's test can be applied. In this case, it would be wise to supplement the test with data on the relative frequency of insertion, deletion and substitution errors of each algorithm to gain more insight into the relative performance. However, this proposition is clearly unworkable for recognition of connected word sentences and a different test procedure must be developed for this case.

4 A MATCHED-PAIRS TEST

Let us suppose that we can divide the output stream from a speech recognition algorithm into segments in such a way that the errors in one segment are statistically independent of the errors in any other segment. A natural candidate for such a segment is a phrase (after which the speaker pauses) or a sentence. Let N_1^i be the number of errors made on the i'th segment by A_1 and N_2^i the number of errors made by A_2. Note that the type of error is unimportant, as long as the method of counting errors is consistent for each segment and for both algorithms; for instance, the method described in [5] could be used. Let $Z^i = N_1^i - N_2^i, i = 1, 2, \ldots, n$, where n is the number of segments. Let μ_Z be the unknown average difference in the number of errors in a segment made by the two algorithms. We would like to ascertain whether $\mu_Z = 0$. A natural estimate of μ_Z would be $\hat{\mu}_Z = \sum_{i=1}^{n} Z_i/n$. The estimate of the variance of the Z_i's is:

$$\hat{\sigma}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \hat{\mu}_Z)^2 \tag{14}$$

The estimate of the variance of $\hat{\mu}_Z$ is then:

$$\hat{\sigma}_\mu^2 = \frac{\hat{\sigma}_Z^2}{n} \tag{15}$$

($\hat{\sigma}_\mu$ is sometimes known as the *standard error of the mean*) If W is defined as:

$$W = \frac{\hat{\mu}_Z}{(\hat{\sigma}_Z/\sqrt{n})} \tag{16}$$

then if n is large enough (> 50, say), W will approximately have a Normal distribution with unit variance. The null hypothesis is $\mathbf{H_0}$: $\mu_Z = 0$, so that under $\mathbf{H_0}$, the distribution of W also has zero mean. $\mathbf{H_0}$ is tested according to the methods described in section 2.

In principle, there is no reason why we need confine ourselves to considering phrases or sentences, as long as we can ensure that the errors in a segment are independent of errors in any other segment. For instance, we could divide the algorithm output into segments where no errors have occurred for some minimal time period T ('good' segments) and segments where errors occur ('bad' segments). T must be sufficently long to ensure that after a good segment, the first error in a bad segment is independent of any previous errors. The segments of the two algorithms could be aligned by aligning each in turn to a transcription of the text using a Dynamic Programming

method [2]. In the error analysis, only the 'bad' segments need to be considered.

The test described is sometimes referred to as a *matched pairs* test, because it considers the effects of two different treatments (algorithms) on equivalent subjects (speech segments). Although it is quite general in its applicability, it does depend on having a sufficently large number of segments for the assumption that W is Normally distributed to be reasonable and to make possible a good estimate of the variance of the Z_i's.

5 RESEARCH ISSUES

In this section, we discuss some possible extensions to the tests we have described and point out some other areas of research where a greater awareness of statistical techniques would be beneficial.

- The tests we have presented compare two algorithms, but very often we would like to decide between more than two algorithms or rank the algorithms. There are generalizations to McNemar's test (e.g. Cochran's test, [3]) which could be used to allow comparison of more than two algorithms.

- Throughout this paper, we have only used the hypothesis testing paradigm, but it would be of considerable interest to give a confidence interval for the difference in the two error-rates. A confidence interval would give a range of values for the difference which is consistent (in a certain technical sense) with the observed data. We may reject the null hypothesis with a very small 'P-value', but if the magnitude of the difference in the error-rates is very small, the improvement we have discerned may be immaterial.

- Currently, algorithms are often based on a rather small amount of training-data and often from a single speaker. It would be of considerable interest and value to be able to make predictions about the change in performance if different training -data or a different speaker were used. There is a vast statistical literature on the design and analysis of experiments that could be profitably explored to study questions such as these.

6 CONCLUSIONS

Two tests for deciding whether the difference in performance of two algorithms tested on the same data-set is significant have been described. McNemar's test is applicable in cases where errors can be considered to be independent (e.g. isolated word recognition with no language model, recognition of short connected word phrases), the matched-pairs test in cases where the algorithms's output can be divided into segments in which the errors are independent of errors in other segments. There is a large statistical literature that can fruitfully be applied to problems of experimental design and analysis in speech science.

ACKNOWLEDGEMENTS

Some of the ideas in this paper arose in conversations with Bob Roth and Jim Baker at Dragon Systems. Acknowledgment is made to the Director of Research, British Telecom Research Laboratories for permission to publish this paper.

References

[1] P. Armitage. *Statistical Methods in Medical Research.* Blackwell Scientific Publications, 1971.

[2] M.J. Hunt. Evaluating the performance of connected word speech recognition systems. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 457–460, April 1988.

[3] E.L. Lehmann. *Nonparametrics.* Holden Day, 1975.

[4] I. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.

[5] D.S. Pallett. Test procedures for the March 1987 DARPA benchmark tests. In *Proc. DARPA Speech Recognition Workshop, San Diego*, March 1987.