IDIAP COMMUNICATION

AN INFORMATION THEORETIC MEASURE OF SEQUENCE RECOGNITION PERFORMANCE

Andrew C. Morris

IDIAP-COM 02-03

April 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-COM 02-03

AN INFORMATION THEORETIC MEASURE OF SEQUENCE RECOGNITION PERFORMANCE

Andrew C. Morris

April 2002

Abstract

Sequence recognition performance is often summarised first in terms of the number of hits (H), substitutions (S), deletions (D) and insertions (I), and then as a single statistic by the "word error rate" WER = 100(S+D+I)/(H+S+D). While in common use, WER has two disadvantages as a performance measure. One is that it has no upper bound, so it doesn't tell you how good a system is, only that one is better than another. The other is that it is not D/I symmetric, although deletions and insertions are equally disadvantageous. At low error rates these limitations can be ignored. However, for the high error rates which can occur during tests for speech recognition in noise the WER measure starts to misbehave, giving far more weight to insertions than to deletions and regularly "exceeding 100%". Here we derive an alternative summary statistic for sequence recognition accuracy: WIP = H^2/(H+S+D)(H+S+I). The WIP (word information preserved) measure results from an approximation to the proportion of the information about the true sequence which is preserved in the recognised sequence. It has comparable simplicity to WER but neither of its disadvantages.

Keywords: word error rate, mutual information, contingency tables, likelihood ratio

Acknowledgements: This work was carried out within the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project. Thanks to Mathiew Magimai Doss for checking the maths.

Notation

t_{ij}	number of times $class(x, y) = (i, j)$
r_i	$\sum_j t_{ij}$, number of times $x = i$
s_j	$\sum_i t_{ij}$, number of times $y = j$
N	$\sum_{ij} t_{ij}$, sum of counts over whole table
p_i	r_i/N, maximum likelihood estimate for $P(x = i)$
q_j	s_j/N, maximum likelihood estimate for $P(y = j)$
p_{ij}	t_{ij}/N, maximum likelihood estimate for $P(class(x, y) = (i, j))$

1. Introduction

Sequence recognition performance is often summarised first in terms of the number of hits (H), substitutions (S), deletions (D) and insertions (I), and then as a single statistic by the "word accuracy" or "word error rate"

$$WAC = 100\frac{(H-I)}{(H+S+D)},\ WER = (100-WAC) = 100\frac{(S+D+I)}{(H+S+D)} \tag{1}$$

WER has two disadvantages as a performance measure. One is that it has no upper bound, so it doesn't tell you how good a system is, only that one is better than another. The other is that it is not D/I symmetric, although deletions and insertions are equally disadvantageous. At low error rates these limitations can be ignored. However, for the high error rates which can occur during tests for speech recognition in noise [3] WER gives far more weight to insertions than to deletions and regularly "exceeds 100%"(see Appendix C).

Here we derive an alternative summary statistic for sequence recognition accuracy which we call WIP/WIL (percentage of word information preserved/lost)

$$WIP = 100\frac{H^2}{(H+S+D)(H+S+I)},\ WIL = (100-WIP) \tag{2}$$

This measure results from an approximation to the proportion of the information about the true sequence which is preserved in the recognised sequence. It has comparable simplicity to WER but neither of its disadvantages.

We first show that the mutual information (MI) between the true and recognised sequences is proportional to Pearson's statistic used for testing for dependence between true and recognised sequences from the evidence in a confusion matrix. We then derive the WIP score as an approximation to the normalised MI which is suitable for use when only the confusion matrix summary statistics (H, D, S, I) are available.

2. Relation between mutual information and Pearson's large sample statistic

If we are given an $m \times n$ "contingency table" or "cross tabulation" of co-occurrence counts t_{ij} between two sets X and Y of m and n classes, then we can test for statistical dependence or "association" between X and Y by first evaluating Pearson's statistic [4] $L(X, Y)$ (Eq.3)

$$L(X, Y) = \sum_{i,j}\frac{(Ob_{ij}-Ex_{ij})^2}{Ex_{ij}} = \sum_{i,j}\left(t_{ij}-\frac{r_is_j}{N}\right)^2\Big/\left(\frac{r_is_j}{N}\right) = N\sum_{i,j}\frac{(p_{ij}-p_iq_j)^2}{p_iq_j} \tag{3}$$

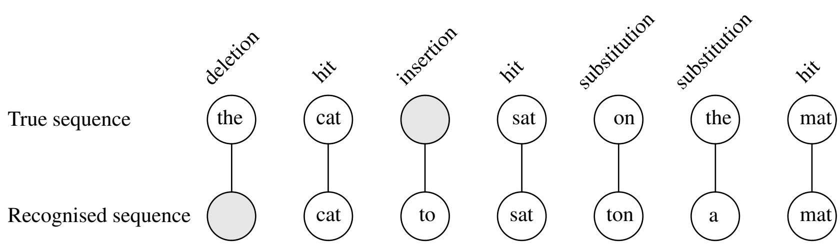

Figure 1. A recognition problem in which insertions and deletions can occur can be converted to a one-one classification by introducing "insertion" and "deletion" as special true and recognised classes.

and then inferring dependence at confidence level ε if $L(X, Y) < \chi^2_{(m-1)(n-1),\,\varepsilon}$. Ob_{ij} in Eq.3 is the number of observed occurrences of $(x, y) = (i, j)$, and Ex_{ij} is the expected number of times this would occur if classes in X and Y were independent, but had the observed occurrence counts.

The Chi-squared test can be applied only to classification problems where each observation must fall into precisely one class. Sequence recognition can be converted into a one-one classification task by considering "insertion" and "deletion" as classes in their own right (Fig.1), and constructing an extended confusion matrix, as in Fig.2.

There is a direct relation between the mutual information $MI(X, Y)$ and Pearson's statistic $L(X, Y)$. If we assume that the estimated probabilities in Eq.3 are accurate estimates of the true probabilities, we can then proceed as

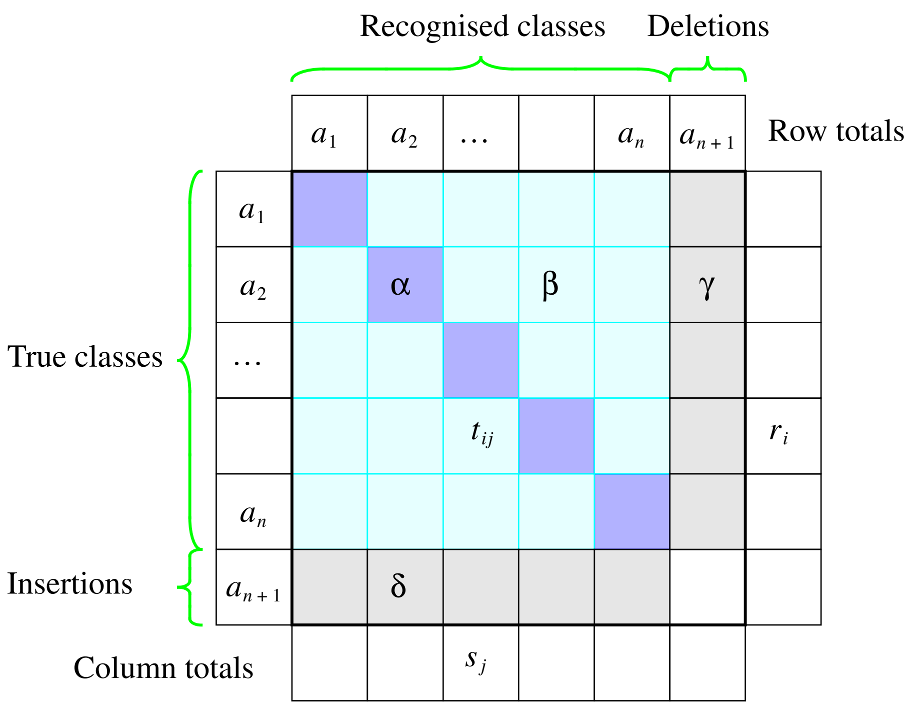

Figure 2. We wish to process a confusion matrix as a contingency table. For this it is necessary to construct a one-one association between true and recognised words. We do this by adding true-word class "insertion" and recognised-word class "deletion" ("word X deleted" becomes "word X recognised as "deletion"). When counts t_{ij} are replaced by summary counts H, S, D, I, we can estimate the original counts by assuming equal counts within each of the areas denoted $\alpha, \beta, \gamma, \delta$, which must sum respectively to H, S, D, I. Estimated row and column sums can then be calculated accordingly.

follows. It is a known result [5] that Pearson's large sample statistic is an *approximation* to the likelihood ratio, λ, used for testing the hypothesis that class sets X and Y are independent (Appendix A).

$$L(X, Y) \cong -2Log_e\lambda \text{, where } \lambda = \prod_{i,j}\left(\frac{p_i q_j}{p_{ij}}\right)^{Np_{ij}} \tag{4}$$

Mutual information can also be expressed in terms of the likelihood ratio statistic [1] (see Appendix B).

$$MI(X, Y) = \sum_{i,j} p_{ij} Log_2 \frac{p_{ij}}{p_i q_j} = \frac{-1}{NLog_e 2} Log_e \lambda \tag{5}$$

It follows (providing the probabilities in (3) are accurate) that MI can be approximated from Pearson's statistic [2]

$$MI(X, Y) \cong \frac{1}{2NLog_e 2} L(X, Y) = \frac{1}{2NLog_e 2} \sum_{i,j} \left(t_{ij} - \frac{r_i s_j}{N}\right)^2 \Big/ \left(\frac{r_i s_j}{N}\right) \tag{6}$$

2.1 Derivation to an approximation to the mutual information between true and recognised sequences from confusion summary counts H, S, D, I

Notation

$H = \sum_{i=1}^{n} t_{ii}$, $S = \sum_{i=1}^{n} \sum_{j \neq i}^{n} t_{ij}$, $D = \sum_{i=1}^{n} t_{i,n+1}$, $I = \sum_{j=1}^{n} t_{n+1,j}$

$N = H + S + D + I$	total number of hits, substitutions, deletions and insertions
$N_T = H + S + D$	true number of words
$N_R = H + S + I$	recognised number of words
n	number of classes (words in dictionary)

We wish to estimate the mutual information between true and recognised word sequences, using Eq.6. If confusion counts have been replaced by summary counts H, D, S, I, we can approximately reconstruct the count values t_{ij} by assuming equal values $t_\alpha, t_\beta, t_\gamma, t_\delta$ within each of the areas $\alpha, \beta, \gamma, \delta$ in Fig.2. This gives

$$t_\alpha = \frac{H}{n},\ t_\beta = \frac{S}{n(n-1)},\ t_\gamma = \frac{D}{n},\ t_\delta = \frac{I}{n} \tag{7}$$

$$r_{i=1\ldots n} = t_\alpha + (n-1)t_\beta + t_\gamma = \frac{H+S+D}{n} = \frac{N_T}{n},\ r_{n+1} = I \tag{8}$$

$$s_{j=1\ldots n} = t_\alpha + (n-1)t_\beta + t_\delta = \frac{H+S+I}{n} = \frac{N_R}{n},\ s_{n+1} = D \tag{9}$$

With $L(X, Y) = \sum_{i,j} \theta_{ij}$, where $\theta_{ij} = (t_{ij} - r_i s_j/N)^2/(r_i s_j/N)$, we can proceed to evaluate the MI estimate in Eq.6 by dividing $L(X, Y)$ into components (A, B, C, D) corresponding to areas $(\alpha, \beta, \gamma, \delta)$ in Fig.2 as follows.

$$\text{A:} \quad \sum_{i,j \in \alpha} \theta_{ij} = n\left(\frac{H}{n} - \frac{N_T N_R}{n^2 N}\right)^2 \Big/ \frac{N_T N_R}{n^2 N} \cong \frac{nH^2 N}{N_T N_R} \tag{10}$$

$$\text{B:} \quad \sum_{i,j \in \beta} \theta_{ij} = n(n-1)\left(\frac{S}{n(n-1)} - \frac{N_T N_R}{n^2 N}\right)^2 \Big/ \frac{N_T N_R}{n^2 N} \cong \left(\frac{NS^2}{N_T N_R} - 2S + \frac{N_T N_R}{N}\right) \tag{11}$$

C: $$\sum_{i,\,j\,\in\,\gamma}\theta_{ij} = n\left(\frac{D}{n}-\frac{N_TD}{nN}\right)^2\Big/\frac{N_TD}{nN} = D\left(\frac{N}{N_T}-2+\frac{N_T}{N}\right) \tag{12}$$

D: $$\sum_{i,\,j\,\in\,\delta}\theta_{ij} = n\left(\frac{I}{n}-\frac{N_RI}{nN}\right)^2\Big/\frac{N_RI}{nN} = I\left(\frac{N}{N_R}-2+\frac{N_R}{N}\right) \tag{13}$$

The term A is larger than terms B, C and D by a factor n, so $MI(X, Y) \propto (A + B + C + D)/N \cong nH^2/N_TN_R$. Dividing by its maximum value (n when $S = D = I = 0$) we obtain the normalised MI approximation which we call the "proportion of information preserved" value.

$$WIP = \frac{H^2}{N_TN_R} = \frac{H^2}{(H+S+D)(H+S+I)} \tag{14}$$

This value is D/I symmetric. It is also strictly increasing in H and decreasing in S, D and I, having a maximum value of 1 when $S = D = I = 0$ and a minimum value of 0 when either $H = 0$ or S, D or I tend to infinity. If we do not normalise, and multiply by $2NLog_e2$, then we obtain $2nN(Log_e2)H^2/N_TN_R$ as an estimate for Pearson's statistic L, from which (if required) we could infer at confidence level ε that the recogniser tells us nothing at all about the word sequence if $L < \chi^2_{n^2,\,\varepsilon}$.

3. Conclusion

We have introduced a new summary statistic for sequence recognition performance, which we call WIP (word information preserved). This was derived as an approximation to the proportion of information which a sequence recogniser preserves about the true sequence.

Like the standard WER (word error rate) measure, WIP is a simple function of the H, D, S, I counts which are often used to summarise a recognition confusion matrix. However, WIP also has the following three desirable properties which he standard WER measure does not have.

- it is a true percentage
- it is D/I symmetric
- it directly approximates the proportion of information preserved

The standard WER score is very well established and cannot be expected to be displaced by any new measure overnight. Unfortunately there is little point in using a performance measure which most other people cannot relate their own results to. However, for work in areas which typically involve high error rates, such as speech recognition in noise, the standard WER measure is particularly unsuitable because (as a quick look at the WER derivatives with respect to D and I will clearly show) the higher the error rate, the more significance it gives to insertions over deletions, so the more inaccurate it becomes (see Fig.3 in Appendix C). Other application areas in which a more accurate performance measure would be beneficial would include iterative training procedures in which a performance estimate is used at each iteration to tune some of the recognition model parameters, and experiments comparing human with machine recognition performance, where absolute as well as relative scores are of interest.

It is clear that a lot of approximations were made in the derivation of this measure. A more accurate estimate of the WIP value can be obtained, when required, by evaluating the mutual information estimate (Eq. 6) directly from the original confusion table counts instead of from the H, S, D, I summary statistics.

Appendix A: Pearson's statistic is an approximation to the logarithm of the likelihood ratio statistic which is used for testing the hypothesis that two sets of classes are independent

We show that $L(X, Y)$ is an approximation to $2NLog_e\lambda$, where $\lambda = \prod_{i,j}\left(\frac{p_i q_j}{p_{ij}}\right)^{Np_{ij}}$ is the likelihood ratio statistic.

$$Log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} \ldots \text{ for } x > -1 \tag{15}$$

$$Log_e\lambda = Log_e\prod_{i,j}\left(\frac{p_i q_j}{p_{ij}}\right)^{Np_{ij}} = -N\sum_{i,j} p_{ij} Log_e\left(\frac{p_{ij}}{p_i q_j}\right) = -N\sum_{i,j} p_{ij} Log_e\left(1 + \frac{p_{ij} - p_i q_j}{p_i q_j}\right) \tag{16}$$

$$= -N\sum_{i,j}(p_{ij} - p_i q_j + p_i q_j)\left[\left(\frac{p_{ij} - p_i q_j}{p_i q_j}\right) - \frac{1}{2}\left(\frac{p_{ij} - p_i q_j}{p_i q_j}\right)^2 + \frac{1}{3}\left(\frac{p_{ij} - p_i q_j}{p_i q_j}\right)^3 \ldots\right] \tag{17}$$

$$= -N\sum_{i,j}\left[(p_{ij} - p_i q_j) + \frac{(p_{ij} - p_i q_j)^2}{p_i q_j} - \frac{1}{2}\frac{(p_{ij} - p_i q_j)^3}{(p_i q_j)^2} - \frac{1}{2}\frac{(p_{ij} - p_i q_j)^2}{p_i q_j} + \ldots\right] \tag{18}$$

$$= -N\sum_{i,j}\left[(p_{ij} - p_i q_j) + \frac{1}{2}\frac{(p_{ij} - p_i q_j)^2}{p_i q_j} - \frac{1}{6}\frac{(p_{ij} - p_i q_j)^3}{(p_i q_j)^2} + \ldots\right] \tag{19}$$

But $\sum_{i,j} p_{ij} = 1$ and $\sum_{i,j} p_i q_j = \sum_i p_i \sum_j q_j = \sum_i p_i = 1$, so $\sum_{i,j}(p_{ij} - p_i q_j) = 0$. Also, $\left|\frac{p_{ij} - p_i q_j}{p_i q_j}\right| < 1$, so higher powers can be ignored. Therefore

$$Log_e\lambda \cong -\frac{1}{2}N\sum_{i,j}\frac{(p_{ij} - p_i q_j)^2}{p_i q_j} = -\frac{1}{2}L(X, Y) \tag{20}$$

Appendix B: Discrete mutual information is directly proportional to the logarithm of the likelihood ratio statistic

$MI(X, Y)$ is proportional to $-Log_e\lambda$, where λ is the likelihood ratio $\prod_{i,j}(p_i q_j / p_{ij})^{p_{ij}}$ used for testing the hypothesis that X and Y are independent.

$$MI(X, Y) = \sum_{i,j} p_{ij} Log_2 \frac{p_{ij}}{p_i q_j} = -Log_2\prod_{i,j}\left(\frac{p_i q_j}{p_{ij}}\right)^{p_{ij}} \tag{21}$$

$$= -\frac{1}{NLog_e 2} Log_e\prod_{i,j}\left(\frac{p_i q_j}{p_{ij}}\right)^{Np_{ij}} = -\frac{1}{NLog_e 2} Log_e\lambda \tag{22}$$

Appendix C: Comparison of WIP and WER scores in a typical experiment in noise robust speech recognition

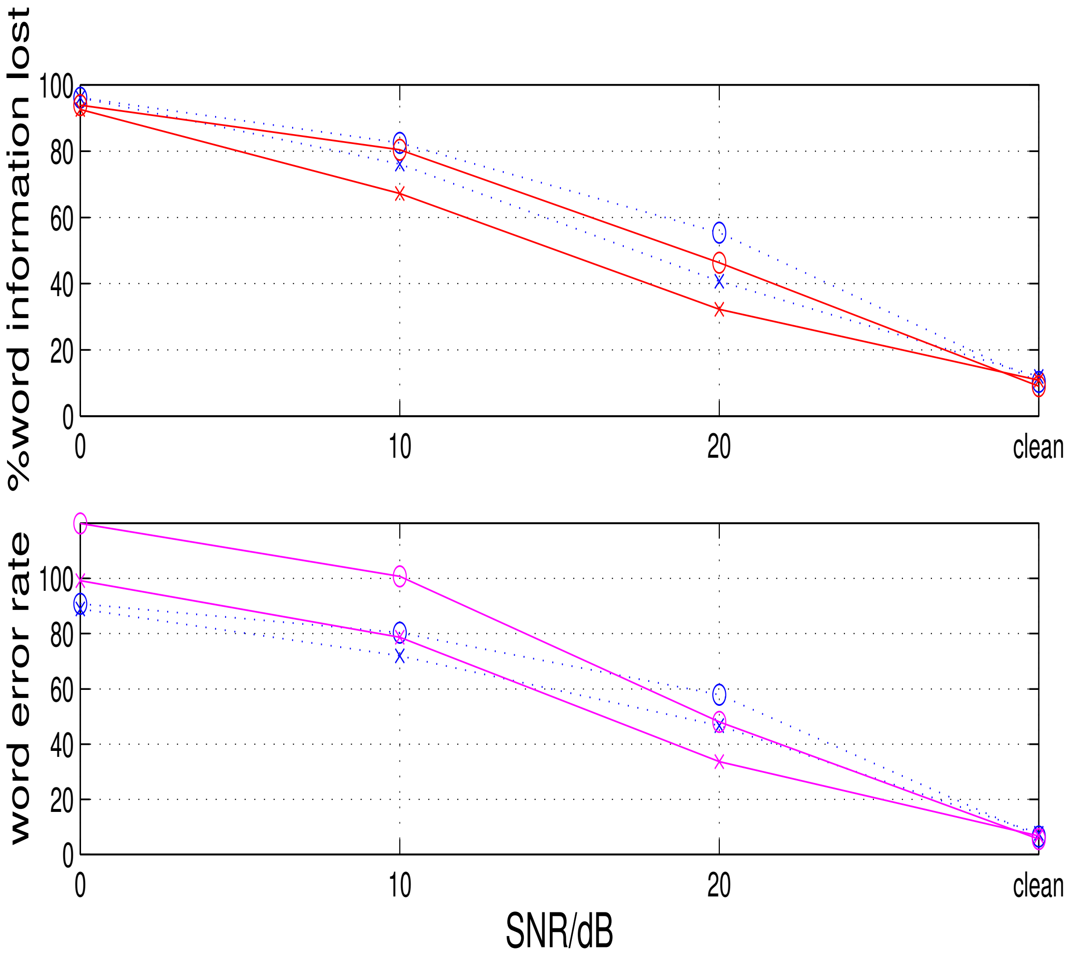

Figure 3. Figure (taken from [3]) compares %WIP and WER scores for a typical experiment in noise robust speech recognition. Upper figure shows %WIP scores for two different noise types (circles = subway, crosses = babble) for speaker independent connected digits recognition using implicit (dotted) and explicit (solid) state duration models. Lower figure shows corresponding WER scores for same experiment.

Figure 3 shows clearly how use of the different WIP and WER performance measures can completely change the apparent performance characteristics of a speech recognition system under different noise conditions. One could complain that the use of WIP in preference to WER in this case would be “simply selecting the performance measure which gives the preferred recognition results”. However, this criticism would be unjustified because the analysis presented in this report has demonstrated that the WIP score has a firm theoretical basis, while the WER score (introduced on purely intuitive grounds) is subject to the theoretical disadvantages which have been mentioned.

References

[1] Miller, G.A. (1954) “Note on the bias of information estimates”, in **Information Theory and Psychology**, (H. Quastler, ed.), The Free Press, Glencoe, IL, pp.95-100.

[2] Morris, A.C. (1992) “An information-theoretical study of speech processing in the peripheral auditory system and cochlear nucleus: application to the recognition of French voices stop consonants”, PhD thesis, Institut de la Communication Parlée, INPG, Grenoble, France.

[3] Morris, A.C., Payne, S. & Bourlard, H. (2001) “Low cost duration modelling for noise robust speech recognition”, Proc. ICSLP 2002 (in press).

[4] Neave, H.R. (1979) **Elementary Statistical Tables**, George Allen & Unwin Ltd, London.

[5] Open University (1977) **Hypothesis testing**, course M341, Unit 11.