From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition

Andrew C. Morris[1], Viktoria Maier[2] & Phil Green[3]

Institute of Phonetics
Saarland University, Germany

morris at coli.uni-saarland.de, maier at idiap.ch, p.green at dcs.shef.ac.uk

[1]Saarland University, Germany; [2]SpandH, Sheffield University (now at IDIAP, Switzerland); [3]SpandH, Sheffield University, UK

Abstract

The word error rate (WER), commonly used in ASR assessment, measures the cost of restoring the output word sequence to the original input sequence. However, for most CSR applications apart from dictation machines a more meaningful performance measure would be given by the proportion of information communicated. In this article we introduce two new absolute CSR performance measures: MER (match error rate) and WIL (word information lost). MER is the proportion of I/O word matches which are errors. WIL is a simple approximation to the proportion of word information lost which overcomes the problems associated with the RIL (relative information lost) measure that was proposed half a century ago. Issues relating to ideal performance measurement are discussed and the commonly used Viterbi input/output alignment procedure, with zero weight for hits and equal weight for substitutions, deletions and insertions, is shown to be optimal.

1. Introduction

Automatic test procedures are necessary for rapid system development, although the ultimate test is always with field trials and the aim of automatic tests should be to emulate human judgement. A single summary statistic is required to permit graphical comparison of system performance, and this should have an intuitive interpretation. The present de facto standard index for ASR system assessment is the Word Error Rate, which is defined as the proportion of word errors to words processed. Let *H, S, D* and *I* denote the total number of word hits, substitutions, deletions and insertions (see Fig.1). *Let* N_1, N_2 and *N* denote the total number of input words, output words, and matched I/O word pairs. For Isolated Word Recognition (IWR) WER is defined [10] as

$$WER\,(IWR) = \frac{S}{N = H + S} = 1 - \frac{H}{N} \qquad (1)$$

In this case there is no problem as the proportion of errors to words processed is an absolute value with a clear interpretation as the probability of incorrect word recognition, or the expected proportion of words incorrect. In Connected Speech Recognition (CSR) some kind of alignment procedure is normally used to pair off input/output words and so decide which pairs should count towards the H, S, D and I counts. Viterbi alignment is usually applied for this purpose, so as to minimise the total error count *(S+D+I)*. WER in CSR is then defined as the ratio of the number of errors to the number of words input [1][7].

$$WER\,(CSR) = \frac{S + D + I}{N_1 = H + S + D} \qquad (2)$$

Unlike the definition of WER for IWR, this it is not I/O symmetric, and it is easily shown that (under normal conditions) it has an upper bound not of 1 but of $\max(N_1,N_2)/N_1$.

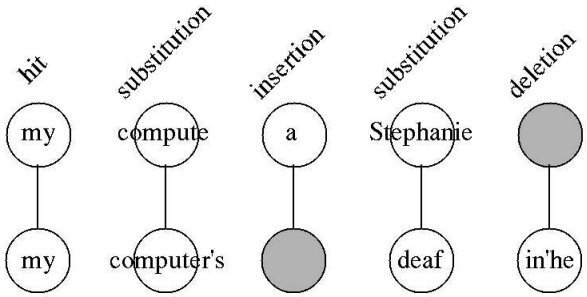

Fig.1 CSR recognition can be converted to IWR-like 1-1 classification by introducing "insertion" and "deletion" words (input words "my computer's deaf in'he?").

For a dictation machine the objective is to minimise the cost of typing in corrections, which could be infinite, and the definition (2) of WER in terms of edit cost is appropriate. However, for most applications the objective is rather to communicate, and correction by typing is not possible, so the hypothetical edit cost is not meaningful. Consider an ASR system for telephone information retrieval which outputs a wrong word for each word input, so that $S=N_1$, $H=D=I=0$ and $WER = N_1/N_1 = 100\%$. A similar system which outputs two wrong words for each input word has $S=N_1$, $H=D=0$, $I = N_1$, and $WER = (N_1+N_1)/N_1 = 200\%$. However, both systems communicate exactly *zero* information and have zero utility, so the performance difference suggested by relative WER scores is, in this case, simply misleading.

What would the ideal ASR performance measure be like? On the one hand the cost of different types of error are application dependent, but on the other, some kind of universal measure is required so that the performance of different ASR systems can be directly compared. In Section 2 we argue the

case for certain constraints on ASR performance metric design. In Section 3 we show that these constraints also *uniquely* select the Viterbi procedure for I/O word sequence alignment. A simple method is also introduced whereby detecting isolated errors can be favoured over detecting runs of errors, if required. In Section 4 we briefly introduce the MER performance measure [4]. In Section 5 we discuss the limitations of the previously introduced RIL information based performance measure [6][10]. In Section 6 we show how the new WIL measure [8] overcomes all of these limitations. We end with a discussion and conclusion.

2. What should an ideal ASR performance index measure?

On the one hand an ideal ASR performance index should reflect human judgement, which will depend in turn on the ASR application. On the other, an acceptable metric for ASR system evaluation must be simple to apply and not be language or application dependent. An intuitively appealing measure would be the proportion of (Shannon) information communicated, but this would depend on the model used for encoding, and natural language encoding makes strong use of word context, language specific grammatical structure and complex pragmatics. If we are to measure information, then we must therefore settle for a model of speech encoding which is context free, i.e. in which all that is being communicated *is a series of independent words*. This gives us our first constraint, which justifies basing our measure on word level HSDI counts:

Measure only "word level" information. (3)

All of the performance measures discussed in this article are based on HSDI counts. It could be objected that the I/O word alignment which these counts require will depend on the arbitrary choice of alignment procedure. However, we show in Section 3 that a small number of reasonable constraints are sufficient to uniquely specify the alignment procedure. Our second constraint expresses the fact that, in most cases, human recognition will identify all of the correct words present.

Maximise the number of hits. (4)

Word level time alignment data for both input and output utterances is often available and could be used to give a more accurate error score *by disallowing alignment between words with no time overlap and forcing alignment, where possible, between words which do overlap*. Such an alignment procedure would be insensitive to the relative weights given to *S,D,I*. Even when (for whatever reason) time alignment data is ignored, we must therefore insist that substitutions are always forced where possible in favour of deletion/insertion pairs (hit, ins., del., hit → hit, sub., hit).

Maximise substitutions before deletions or insertions (5)

In the next section we show that the above constraints alone are sufficient to identify standard Viterbi alignment, using equal *SDI* error weights, as optimal. This means that the *HSDI* counts, on which the information based ASR scores introduced in later sections depend, are uniquely determined. It also means that any use of non equal *S, D* and *I* weights (such as referred to in [11]) requires special justification.

3. Optimal input/output alignment

For the purpose of word-level recognition performance it is first necessary to obtain *HSDI* counts. This is normally achieved by using Viterbi search to efficiently select the I/O word sequence alignment for which the weighted error score is globally minimised. But which weights should be used?

3.1. Optimal HSDI weighting for I/O alignment

The linear loss function usually used with Viterbi can assign a different "loss" weight to each error type,

$$loss = H W_H + S W_S + D W_D + I W_S \quad (6)$$

Minimising this objective is equivalent to maximising the probability of the I/O *HSDI* match label sequence when match type occurrences are independent and the above weights give the negative log probability of each match type occurring. However, *H, S, D & I* are not independent, because $N_1=H+S+D$ and $N_2=H+S+I$. There are therefore not more than 2 free weights. A key point here is that *global minimisation of this linear objective will always maximise the count with minimum weight first*, then the count with the next smallest weight, and so on. By (4) H must be maximised first, so we can set $W_H=0$ and all other weights > 0. Then, as $N1-N2=D-I$, we have $DW_D+IW_I = D(W_D+W_I)+constant$, so minimisation is not affected if W_D+W_I is replaced by a single weight, W_E. This leaves only SW_S+DW_E, but minimisation is unchanged by applying any scaling factor to the objective, so the only choice we have is to decide whether W_S is greater or less than W_E. However, by (5) *D&I* must be minimised before *S*, so W_E must be $> W_S$. *Only linear weightings with* $W_H<W_S<W_E$ *are therefore permissible, and all such weightings are equivalent.* The usual *min(S+D+I)* objective fulfills this condition. Therefore, though the relative cost of different types of error may be application dependent, the standard zero *H* and equal *S, D* and *I* error weights is optimal.

3.2. Weighting between isolated occurrences and runs

In some applications we may wish to downweight errors which are more easily detected and possibly corrected, such as isolated as opposed to contiguous errors. Weighting between isolated occurrences and runs of any of the *HSDI* match types can be effected by splitting the count for the match type concerned (e.g. *D*) into counts D_1 and D_2, according to whether the preceding match type was different or equal, respectively. For example, to maximise first *H* then *S* then D_1 (possibly isolated deletions) then D_2 (deletions occurring in runs), we would minimise $S+D_1+2D_2+I$ (or equivalently, $S+2D_1+3D_2$, as $D=D_1+D_2$ and minimising *D* is equivalent to minimising $I=D+N2-N1$). Importantly, Viterbi is still guaranteed to globally optimise any monotonic function of such counts, as they are Markovian. In terms of a maximum probability objective, this weighting models run length as having a Geometric pdf, $P(len=d) = ß(1-ß)^{1-d}$, with $ß = P(len=1) = 1/(average\ run\ length)$. If in the context *"P(Di)"* D_i denotes the event that a *D* is of type *Di (i=1 or 2)*, then *ß* is related to W_{D1} and W_{D2} *by* $ß/(1-ß) = P(D_1)/P(D_2) = exp(W_{D2}-W_{D1})$. By (5), W_S must be $<W_{D1}+W_I$. The (posterior) average run length (of deletions) can be found from D_1 and D_2 as $(D_1+D_2)/D_1$.

3.3. Summing over all possible alignments

An alternative form of *HSDI* estimation has been investigated which avoids imposing any rules for I/O alignment [4]. In this approach counts are obtained as a weighted sum of counts over all possible alignments. However, while this can be done (using the forward-backward procedure sometimes used instead of Viterbi for decoding in ASR [3]), if this approach is not to violate (4) or (5), which we have shown here to select the Viterbi solution as unique, then (except in the rare case of co-optimal alignments) the weight (sequence probability) for any alignment not selected by Viterbi must be zero, so summation would be a waste of time.

4. Match error rate (MER)

If all that was required was to obtain an absolute measure of CSR performance, one could simply divide WER by its maximum possible value, $max(N_1,N_2)/ N_1$, to obtain

$$normalised\,WER = (S+D+I)/\max(N_1,N_2) \tag{7}$$

However, a more intuitively appealing measure is given by the Match Error Rate (MER) (8). The ranking behaviour of MER is between that of WER and WIL (see Table 1).

$$MER = \frac{S+D+I}{N=H+S+D+I} = 1-\frac{H}{N} \tag{8}$$

MER is the probability of a given match being incorrect. In Sections 5 and 6 we introduce information theoretic measures of word information communicated. These depend on word confusion probabilities which are generated by I/O alignment, which we have now shown has a good foundation.

5. Relative information lost (RIL)

The Relative Information Lost *(RIL)* (12) was proposed as a measure of ASR system performance half a century ago [5][10][4]. Mutual information *(I, or MI)* [9] here provides a measure of the statistical dependence between the input words *X* and output words *Y* in the unordered set of I/O word pairs obtained by I/O alignment. *MI* is the fall in uncertainty about *Y* when given *X*, or vice versa (11). H and MI are $>= 0$, so $I(X,Y) <= min(H(X),H(Y))$. The mapping from X to Y here is deterministic, so $H(Y) <= H(X)$. Therefore $0 <= I(X,Y) <= H(Y)$. Both *RIL* and its complement, $RIP = I(X,Y)/H(Y)$, are therefore $>= 0$ *and* $<= 1$. *RIP* measures the proportion of word information preserved in recognition. *MI* is defined in terms of (Shannon) entropy *H*, the expected log inverse probability of a random variable. For discrete X with values $x_1..x_n$, occurring with probabilities $P(x_i)$, $H(X)$ is given by (9).

$$H(X) = E[\log 1/P(X)] = -\sum_{i=1}^{n} P(x_i)\log P(x_i) \tag{9}$$

$$H(Y|X) = -\sum_{i,j} P(x_i,y_j)\log P(y_i|x_i) \tag{10}$$

$$I(X,Y) = H(X)-H(X|Y) = H(Y)-H(Y|X) = \sum_{i,j} P(x_i,y_j)\log\left[P(x_i,y_j)/P(x_i)P(y_j)\right] \tag{11}$$

$$RIL = 1-\frac{I(X,Y)}{H(Y)} = \frac{H(Y|X)}{H(Y)} \tag{12}$$

The probabilities in (9, 10, 11) can be (crudely) estimated from relative frequencies in the matrix of I/O word confusion counts C_{ij} (augmented by del. and ins. "word" counts),

$$N = \sum_{ij} C_{ij}, \quad P(x_i,y_j) \cong C_{ij}/N$$
$$P(x) \cong \sum_j C_{ij}/N, \quad P(y_j) \cong \sum_i C_{ij}/N \tag{13}$$

The RIL measure was not widely taken up. The two main reasons for this are probably that it is not as simple to apply as WER ((2) vs. (9, 10, 11)), and it measures zero error for ***any*** one-one mapping between input and output words, which is unacceptable [4]. In Section 6 we review the recently introduced WIL measure which is closely related to RIL, but overcomes both of these limitations.

6. Word information lost (WIL)

An approximation to RIL which is based on HSDI counts alone, and does not use logarithms, was derived in [8]. This was done by reconstructing a confusion matrix from HSDI counts, then making use of a little known relation between MI and Pearson's Large Sample Statistic L(X,Y) (as used in Chi-squared tests for independence). MI is linked to L in [8] via the likelihood ratio statistic, λ (14), for which [2] gives (15) and [5] gives (16),

$$\lambda = \prod_{i,j}\left(\frac{P(x_i)P(y_j)}{P(x_i,y_j)}\right)^{NP(x_i,y_j)} \tag{14}$$

$$L(X,Y) = N\sum_{i,j}\left[P(x_i,y_j)-P(x_i)P(y_j)\right]^2/P(x_i)P(y_j) \cong -2\log_e\lambda \tag{15}$$

$$I(X,Y) \cong -\log_2\lambda/N \tag{16}$$

from which it follows that

$$I(X,Y) \cong L(X,Y)/2N\log_e 2 \tag{17}$$

In [8] it is shown that when the MI approximation given by (17) is applied to the confusion matrix reconstructed from *HSDI* counts, the result is dominated by contributions from the leading diagonal "hit" entries,

$$diag = N\sum_{i=1}^{n}\left[P(x_i,y_i)-P(x_i)P(y_i)\right]^2/P(x_i)P(y_i) \tag{18}$$

The relevant *HSDI* reconstructed probabilities are

$$P(x_i,y_i) \cong H/n, \quad P(x_i) \cong N_1/n, \quad P(y_i) \cong N_2/n \tag{19}$$

Dividing *diag* by its max. poss. value (nN when $H = N$), gives

$$\frac{I(X,Y)}{H(Y)} \cong \frac{(H-N_1N_2/nN)^2}{N_1N_2} \tag{20}$$

Subject to $H >> S+D+I$, this simplifies to give (21) for the definition of WIP "word information preserved" and the proposed "word information lost" measure, *WIL = 1-WIP*.

$$WIP = \frac{H}{N_1}\frac{H}{N_2} \cong \frac{I(X,Y)}{H(Y)}, \quad WIL = 1-WIP \tag{21}$$

Retaining only contributions to L from the leading diagonal of the *HSDI* reconstructed confusion matrix has the important effect not only of leading to a simple formula for *WIL*, but also of creating a form of mapping-restricted *MI (MRMI)* which is sensitive to the required (identity) I/O mapping implicit in the hit count - which *RIL* is not. Although *WIP* is only an accurate measure of the proportion of *MRMI* preserved when N is dominated by *H*, it can also be interpreted simply as the probability that any input word is matched with an equal output word *and* vice versa.

Input	Output	H	S	D	I	%WER	%MER	%WIL
X	X	1	0	0	0	0	0	0
Xiii	XXYY	1	0	0	3	300	75	75
XYX	XZd	1	1	1	0	67	67	83
X	Y	0	1	0	0	100	100	100
Xi	YZ	0	1	0	1	200	100	100

Table 1. WER, MER and WIL do not always give the same ranking (X, Y and Z are arbitrary words, i = insertion, d = deletion)

7. Discussion

The commonly used WER measure is ideally suited only to CSR applications where output errors can be corrected by typing. For almost any other type of speech recognition system a measure based on the proportion of information communicated would be more useful.

We have argued that for practical reasons it is necessary to restrict the information measured to context free word information only (3). Word-information measures are based on confusion matrix counts and to obtain these it is necessary to perform input/output word sequence alignment. We showed that, under the conditions that hits should be maximised first (4), and then substitutions (5), any alignment procedure minimising a linearly weighted sum of hit, substitution, deletion and insertion counts is optimal proving only that $W_H < W_S < W_D+W_I$, and all such procedures are equivalent. It was also shown how weighting between isolated occurrences and runs can be effected, if required, by separating each count type of interest into counts for instances occurring first in a run and otherwise.

The previously proposed information theory based RIL index was then presented and two reasons were identified to explain why it was not taken up as a standard measure by the ASR community: it was much more complicated than WER to use and also had the theoretical disadvantage that it would estimate zero error for any one-one mapping from input to output words (whereas the only mapping of interest is the identity mapping). We then reviewed the recently proposed Word Information Lost (WIL) performance index, which avoids both of these problems. It was shown how WIL was derived as a simple function of HSDI counts by making use of an equation linking mutual information with Pearson's Large Sample Statistic, and then retaining only dominant terms.

8. Conclusion

MER and WIL both provide simple CSR performance measures which vary from 0 when there are no errors to 1 when there are no hits. They are more suitable than WER for the evaluation of any application in which the proportion of word information communicated is more meaningful than edit cost. At low error rates all three give similar scores so that the inappropriate theoretical basis for the WER measure is not noticeable. However, in tests for many real world applications, where significant error rates are common and choosing the best system is very important, the rankings given by each measure start to differ significantly and the risk of incorrect choice of best performing system becomes very real. Statistical tests used with WER can also be applied to MER and WIL. Both MER and WIL have intuitively simple probabilistic interpretations, but in so far as WIL measures the proportion of (mapping sensitive) word information communicated (at least, when N is dominated by H), and communication is mainly what speech is for, it is the preferred measure.

9. Acknowledgements

This work was carried out partly within the EC RESPITE (REcognition of Speech by Partial Information Techniques), HOARSE (Hearing Organisation And Recognition of Speech in Europe) and SecurePhone projects.

10. References

[1] Bahl, L.R. & Jelinek, F. "Decoding for channels with insertions, deletions and substitutions, with applications to speech recognition", IEEE Trans. Inform. Theory, IT 21 No.4, pp.404-411, 1975.

[2] Collings, S.N. *Fundamentals of Statistical Inference, Unit 11. Hypothesis Testing II*, course M341, The Open University Press, pp.105-107, 1977.

[3] Gold, B. & Morgan, N. *Speech and audio signal processing*, Wiley, 2000.

[4] Maier, V. "Evaluating RIL as basis of automatic speech recognition devices and the consequences of using probabilistic string edit distance as input", Univ. of Sheffield, third year project, 2002.

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

[6] Miller, G. & Nicely, P. "An analysis of perceptual confusions among some English consonants", J. Acoust. Soc. Am., Vol.27, No.2, 1955.

[7] Moore, R. "Evaluating speech recognisers", Dept. of Elec. Eng., Univ. of Essex, UK, white paper, 1979.

[8] Morris, A.C. "An information theoretic measure of sequence recognition performance", IDIAP-com 02-03, 2002. ftp://ftp.idiap.ch/pub/reports/2002/com02-03.pdf

[9] Papoulis, A. *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1991.

[10] Woodard, J.P. & Nelson, J.T. "An information theoretic measure of speech recognition performance", Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA, 1982.

[11] Young, S.J. & Chase, L.L. "Speech recognition evaluation: a review of the US CSR and LVCSR programmes", Computer Speech Language, Vol.12, pp.263-279, 1998.