ON THE USE OF THE RELATIVE INFORMATION TRANSMITTED (RIT) MEASURE FOR THE ASSESSMENT OF PERFORMANCE IN THE EVALUATION OF AUTOMATED SPEECH RECOGNITION (ASR) DEVICES

Alan M. Smith

Speech Systems Group
Integrated Systems Laboratory
NYNEX Science and Technology

ABSTRACT - The use of an information-theoretic based metric, the Relative Information Transmitted (*RIT*), may facilitate the assessment of the performance of automated speech recognition (ASR) devices. The *RIT* provides a scalar value which may be employed in a manner similar to the use of such traditional scalar measures as 'percent correct'. The complexity of the recognition task is factored into the computation of the *RIT*. For example, chance-level performance on a two-word recognition task and on a four-word recognition task both yield equivalent *RIT* values of zero, whereas the associated 'percent correct' performance on these tasks would be 50% and 25%, respectively. The *RIT* is an entropy-based characterization of the ASR as a receiver in a communication channel in that the distribution of input/output characteristics of the associated confusion matrix is reflected in the generated measure. However, as with all figures of merit, the use of the *RIT* must be coupled with an understanding of the specific task and application domain associated with the assessment. Examples are provided which indicate that conclusions drawn from the use of the *RIT* may not always be in agreement with those derived from consideration of the 'percent correct' performance of the same system.

INTRODUCTION

A necessary step in the evaluation of an automated speech recognition device (ASR) is the measurement of the objective factors associated with the operating characteristics of the system. At this time, we *assess* the ASR, in the sense of the Webster's (1989) definition, which is *'to determine the importance, size, or value of'*. While both *evaluation* and *assessment* refer to the determination of *worth* or *value*, we consider *assessment* to be associated with the measurement of the objective system characteristics, in a sense *sizing* the system, while *evaluation* is associated with the more colloquial connotation of determining the usefulness or appropriateness of an ASR for an application. This paper considers the utility of a specific measure which might be employed in the assessment process, the *Relative Information Transmitted*, or *RIT*, as presented by Poock (1989) .

As an example of the evaluation process, it is of interest to consider the use of an ASR in an isolated-word, small-vocabulary, speaker-dependent speech recognition task. The outcome of applying the recognizer of interest to the defined task is typically reported in a manner which is a summarized version of the test results. That is, the results of an isolated word recognition test are reported in terms of some single value thought to be descriptive of the overall performance of the system, such as *'percent correct'*, rather than reporting the recognizer output for each instance of every input test word. While resolution is lost with respect to the exact details of the test, the use of such a *figure of merit* provides a mechanism for easily describing the overall performance of the system, in general terms. This, in turn, allows for the comparison of ASRs on the basis of this single measure. Alternatively, replication of a test might be performed using a single ASR, with either systematic variation of system parameters or different databases employed.

RELATIVE INFORMATION TRANSMITTED (*RIT*)

Poock (1989) suggests that an information-theoretic approach which was previously discussed by Woodard (1984) provides an assessment measure which has a different perspective than those often used by speech researchers. In the type of isolated-word recognition task previously outlined, the typical measure of performance used to summarize the results of the test is the *'percent correct'* value of the response of the system to the test inputs. This provides an estimate of the *pro-*

bability of error characteristic of the system. However, this measure provides no information regarding the distribution of the observed errors, with respect to the set of inputs to the system under test. It is therefore suggested that a metric be employed which takes into consideration the performance of the system when taken from the perspective of how much information is input to the system, and how much information is recovered from it.

If we consider the ASR as a communication channel which accepts information as input, and generates information as output, then we can evaluate it in terms of its *information transfer* characteristics. Let $P(\cdot)$ will always denote probability, and let X be a discrete random variable (rv) with possible outcomes $x_1, x_2, ..., x_N$ and *probability mass function*, pmf, $p_X(\cdot)$, i.e.,

$$p_X(x_i) \equiv P(X{=}x_i) > 0 \quad i = 1,2,....,N \quad (1)$$
$$p_X(x_i) \equiv 0 \qquad \textit{otherwise}$$

The *self-information* in a single event X which has a probability of occurrence $P(X)$ is defined by

$$h(X) \equiv -log_2\, P(X) \qquad (2)$$

The *average self-information*, or *entropy*, in the pmf of X (or simply in X) is

$$H(X) \equiv -\sum_{i=1}^{N} p_X(x_i) \log_2 p_X(x_i) \qquad (3)$$

By this definition, $H(X)$ is a weighted average of the self-informations of the events $X = x_i$, $i = 1,2,....,N$.

Suppose that X and Y are *jointly discrete* rvs with joint pmf $p_{XY}(\cdot\,,\cdot)$, where

$$p_{XY}(x_i,y_j) = P(X = x_i, Y = y_j) \qquad (4)$$
$$\text{for } i = 1,2,....,N;\ j = 1,2,.....,M$$
$$p_{XY}(x_i,y_j) = 0 \quad \textit{otherwise}$$

Note that the *marginal*, or individual, pmfs of X and Y can be obtained from the joint pmf. Specifically,

$$p_X(x_i) = \sum_{j=1}^{M} p_{XY}(x_i,y_j) \qquad (5)$$

$$p_Y(y_j) = \sum_{i=1}^{N} p_{XY}(x_i,y_j) \qquad (6)$$

By applying (3) to p_Y and p_{XY}, we obtain the entropies

$$H(Y) = -\sum_{j=1}^{M} p_Y(y_j) \log_2 p_Y(y_j) \qquad (7)$$

$$H(XY) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p_{XY}(x_i,y_j) \log_2 p_{XY}(x_i,y_j) \quad (8)$$

Finally, we define the *average mutual information* in X and Y by the weighted average of the mutual informations in all the event pairs $X = x_i$, $Y = y_j$, or

$$H(X{:}Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} p_{XY}(x_i,y_j) \log_2 \frac{p_{XY}(x_i,y_j)}{p_X(x_i)\, p_Y(y_j)} \quad (9)$$

By applying equations (3) and (5)-(8) to equation (9), it can be seen that

$$H(X{:}Y) = H(X) + H(Y) - H(XY) \qquad (10)$$

When we look at the average mutual information $H(X{:}Y)$, we interpret this as the information about the input to the system which is provided by its output. Thus, if we know or can estimate $H(XY)$, we can assess the ASR in an information-theoretic context. In order that different systems may be compared through the use of this measure of mutual information, we normalize the value of $H(X{:}Y)$ by the entropy of the input to the system. The measure of *information transfer* thus generated is relative to the input to the system, and is denoted the *Relative Information Transmitted*, or *RIT*.

$$RIT = \frac{H(X{:}Y)}{H(X)} \qquad (11)$$

A more complete development of the basis for this approach is given in Pfeiffer (1978) and Woodard (1984). A discussion of the development of the *RIT* may be found in the material of Poock (1989), and reviewed in Smith (1990).

From equations (5), (6), (3), and (7)-(11), we see that to compute *RIT*, it suffices to have the distribution p_{XY}. However, in the convenient tabular representation of the collected test data which we typically use, which is known as a *confusion matrix*, we have the distribution f_{XY}. An example of the prototypical confusion matrix is presented as Table A1 in the Appendix. Each confusion matrix reports the *frequency* with which specific input/output relationships occur in the course of the assessment process. The inputs to the system are represented by the rows of the matrix, while the responses of the system are represented by the columns. The response column labeled *R* is provided to enumerate those inputs to the system which the ASR *rejected*, or did not 'recognize'. This typically means that some parameter value used in the recognition process has (or has not) exceeded a specified threshold, and that the recognition process for that input is aborted. A system exhibiting perfect performance would generate a confusion matrix which contains non-zero values on the main diagonal only of the N-by-N submatrix which does not contain the *R* (rejection) column.

The confusion matrix has been augmented by an auxiliary row and column which contain frequency subtotals used in the estimation of probabilities of occurrence. Subtotals for each row are indicated in the rightmost entry of each matrix row (under *rt*) and subtotals for each matrix column are indicated in the bottom entry of each column (alongside *ct*). The entry at the intersection of the *ct* row and the *rt* column, denoted *TOTAL*, represents the total number of trials in the test represented by the matrix.

We estimate the probabilities needed for the computations in equations (7)-(9) through the use of the following substitutions:

$$p_X(x_i) \approx \rho_X(x_i) \tag{12}$$

$$p_Y(y_j) \approx \rho_Y(y_j) \tag{13}$$

$$p_{XY}(x_i,y_j) \approx \rho_{XY}(x_i,y_j) \tag{14}$$

where

$$\rho_X(x_i) = \frac{f_X(x_i)}{n} \tag{15}$$

$$\rho_Y(y_j) = \frac{f_Y(y_j)}{n} \tag{16}$$

$$\rho_{XY}(x_i,y_j) = \frac{f_{XY}(x_i,y_j)}{n} \tag{17}$$

and n = total inputs.

Statistics associated with the performance of the example systems are represented by the computed values of $H(X)$, $H(Y)$, $H(XY)$, $H(X{:}Y)$, and R/T. In addition, $P(ERR)$ and $P(COR)$ are provided, which are defined as follows:

$$P(ERR) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{XY}(x_i,y_j)_{i \neq j} + \sum_{i=1}^{N} R_i}{TOTAL} \tag{18}$$

$$P(COR) = 1.0 - P(ERR) \tag{19}$$

EXAMPLE DATA

Examples 1-6, inclusive, are exactly those discussed by Poock (1989). Examples 7 and 8 demonstrate system characteristics similar to, but different from, that seen in the data provided by Examples 1 and 6, respectively.

	y_1	y_2	*R*	*rt*
x_1	100	0	0	100
x_2	0	100	0	100
ct	100	100	0	200

Table 1: Example 1

	y_1	y_2	*R*	*rt*
x_1	25	25	0	50
x_2	25	25	0	50
ct	50	50	0	100

Table 2: Example 2

	y_1	y_2	*R*	*rt*
x_1	81	9	0	90
x_2	9	81	0	90
ct	90	90	0	180

Table 3: Example 3

	y_1	y_2	*R*	*rt*
x_1	100	0	0	100
x_2	20	80	0	100
ct	120	80	0	200

Table 4: Example 4

	y_1	y_2	y_3	*R*	*rt*
x_1	40	40	40	0	120
x_2	40	40	40	0	120
x_3	40	40	40	0	120
ct	120	120	120	0	360

Table 5: Example 5

	y_1	y_2	y_3	*R*	*rt*
x_1	180	10	10	0	200
x_2	10	180	10	0	200
x_3	10	10	180	0	200
ct	200	200	200	0	600

Table 6: Example 6

	y_1	y_2	*R*	*rt*
x_1	0	100	0	100
x_2	100	0	0	100
ct	100	100	0	200

Table 7: Example 7

	y_1	y_2	y_3	*R*	*rt*
x_1	10	180	10	0	200
x_2	10	10	180	0	200
x_3	180	10	10	0	200
ct	200	200	200	0	600

Table 8: Example 8

Summary statistics associated with the performance characteristics of the data of these examples are provided in Tables A2 and A3 of the Appendix.

Discussion of Examples

Example 1 shows that when no errors occur, $P(ERR)$ is 0.0 and *RIT* is 1.0, as is expected. There is complete transfer of information through the system.

Example 2 shows that when performance occurs at *chance level*, *RIT* is 0.0. This also is expected as no information has been transmitted through the system.

Example 3 illustrates a system which has $P(ERR)$ equal to 0.1, where the error responses are equally likely in any of the input classes. This system has a 2-item vocabulary. The resultant *RIT* is approximately 0.53.

Example 4 is a system which also has $P(ERR)$ equal to 0.1. The error responses, however, are associated with only one of the two input classes. This locality of error results in a value of *RIT* of approximately 0.61, this higher value reflecting the fact that the errors are generated as responses to only one of the inputs.

Example 5 illustrates a *higher order* system, in that it utilizes a three-item vocabulary, rather than two, as has been discussed to this point. Like the system of Example 2, it demonstrates *chance level* performance. Because it has three input/output classes (excluding *R*), this results in a $P(ERR)$ of approximately 0.67. As *no information* is transmitted through this system, as in Example 2, it also demonstrates a value of *RIT* equal to 0.0.

Example 6 shows a system which operates with $P(ERR)$ equal to 0.1, as do Examples 3 and 4. The value of *RIT* for this system is about 0.64. When compared with Example 3, which also has a value of $P(ERR)$ equal to 0.1, the effects of system complexity (order) are seen in the value of *RIT*. While in the cases of both Example 3 and Example 6 it is true that the probability of error (0.10) is evenly distributed, the higher order of the Example 6 system yields a higher computed value of *RIT*, which is 0.53 for Example 3.

Example 7 illustrates a very *inaccurate* system, yielding a value $P(ERR)$ of 1.0. The differences in the systems of Examples 1 and 7 are clearly not in the *amount* of information which they transmit, it is in the *nature* of the information. In either example, the input to the system is known, with certainty, based on the response of the system. Taken with this perspective, it is reasonable to find that the value of *RIT* for both systems is 1.0. However, the probability that the response given correctly identifies the input to the system could not be more disparate for these two examples.

Example 8 demonstrates a system whose data can be thought of as the system response of Example 6, where the entries of the (sub)matrix have been rotated to the right by one column. This transformation does not change the relative distribution of the data within the two-dimensional coordinate system which the matrix represents, but rather, a shift in the index associated with the entries in one (horizontal, or *row*) dimension. The result of this manipulation is that the $H(X)$, $H(Y)$, $H(XY)$, and resultant *RIT* values are unchanged. The amount of information transmitted by this system is the same as that seen in the system of Example 6, and *RIT* remains 0.64. However, the *accuracy* of the two systems is enormously different. Whereas $P(ERR)$ for Example 6 was 0.1, it takes on a value of 0.95 for the system in Example 8.

ADDITIONAL EXAMPLE DATA

To further investigate the utility of the *RIT* assessment measure, this technique was applied to the results of tests performed using a *typical* ASR. The tabulated summary statistics of these tests, denoted Examples 9-14, inclusive, are presented in Table A4 of the Appendix. Due to limited space availability in this paper, the confusion matrices for these examples are not provided at this time, but may be found in Smith (1990).

The recognition system utilized was a software implementation of a *dynamic time warping* (DTW) -based, speaker-dependent, isolated-word recognizer under development at the Integrated Systems Laboratory (ISL) of NYNEX. Training and test data were collected at the ISL facility under low-noise, high-bandwidth conditions. The test protocol utilized an eleven word vocabulary, digits *one* through *nine*, *zero*, and *oh*. The database was nominally composed of twenty tokens per digit per speaker. Two tokens of each vocabulary item were used for training. The test phase attempted recognition on each of the remaining eighteen tokens of each digit.

Discussion of Additional Examples

Examples 9 and 10 have three and four errors, respectively, out of a total of 198 trials. As a result, the values of $P(ERR)$ for these systems differ by 33%. However, the values of *RIT* associated with these two examples are equivalent.

Examples 11 and 12 both have four errors out of a total of 198 trials, yielding values of $P(ERR)$ which are identical. The different values of *RIT* for these two systems does not directly reflect this similar error rate.

Example 13 has two errors out of a total of 196 trials; Example 14 has seven errors out of 196 trials. $P(ERR)$ for Example 13 is less than one third of that for Example 14, as would be expected. What is not immediately obvious is the finding that the value of *RIT* for Example 14 is *greater* than that for Example 13, indicating that the system of Example 14 transmits more information in spite of its higher error rate.

SUMMARY

The Relative Information Transmitted assessment parameter described by Poock (1989) appears to have many merits which suggest that it should be included as part of the overall evaluation process. This is particularly true for the case of isolated-word ASRs, where the behaviour of this type of system is easily characterized as a confusion matrix. For example, as this measure reflects the complexity of the task, in some sense, it may provide the basis to compare ASR performance across applications which require different sized vocabularies. However, this measure should be used to *supplement*, rather than replace, existing measures such as 'percent correct', as it addresses a useful, but *different*, set of assessment considerations.

It is of note that the *RIT* reflects information about the *pattern* of the errors made by ASRs. Subsequent utilization of this information may assist in the refinement of ASR design and implementation, such as when early identification of error occurrence localized to specific vocabulary items is provided.

REFERENCES

Pfeiffer, P.E. (1978) *Concepts of Probability Theory*, 2nd Rev. Ed., Dover Publications, Inc., New York, New York. USA, pp. 245-258.

Poock, G.K. (1989) "A Different Information Theory Approach for Calculating Woodard's Relative Loss (RIL) When Evaluating Speech Recognizers", *Journal of the American Voice I/O Society*, Vol. 6, pp. 47-60.

Smith, A.M. (1990) "On the Use of the Relative Information Transmitted (RIT) Measure for the Assessment of Performance in the Evaluation of Automated Speech Recognition (ASR) Devices", *NYNEX Technical Memorandum 90-0004*.

Webster's Ninth New Collegiate Dictionary (1989) Meriam-Webster Inc., Springfield, Massachusetts, USA.

Woodard, J.P. (1984) "Analyzing Performance of Speech I/O Devices with Information Theory", *Journal of the American Voice I/O Society*, Vol. 1, pp. 34-37.

ACKNOWLEDGEMENTS

Thanks go to Pat Sullo of Rensselaer Polytechnic Institute, who contributed significantly to the development of this material and provided a knowledgeable perspective to the consideration of this technique. The speech database utilized to generate the results reported in Examples 9 through 14, inclusive, is a multi-speaker, isolated word (digits) database which is available primarily though the extensive efforts of Harry Chang of ISL.

APPENDIX

	y_1	y_2	·	·	y_j	·	·	·	y_N	R	rt
x_1											·
x_2											·
·											·
x_i					$f_{XY}(x_i,y_j)$						$f_X(x_i)$
·											·
·											·
·											·
x_N											·
ct	·	·	·	·	$f_Y(y_j)$	·	·	·	·	·	$TOTAL$

Table A1: Augmented Confusion Matrix f_{XY}

	Example 1	Example 2	Example 3	Example 4
$P(ERR)$	0.000000	0.500000	0.100000	0.100000
$P(COR)$	1.000000	0.500000	0.900000	0.900000
$H(X)$	1.000000	1.000000	1.000000	1.000000
$H(Y)$	1.000000	1.000000	1.000000	0.970951
$H(XY)$	1.000000	2.000000	1.468996	1.360964
$H(X:Y)$	1.000000	0.000000	0.531005	0.609987
RIT	1.000000	0.000000	0.531004	0.609987

Table A2: Summary Statistics: Examples 1-4

	Example 5	Example 6	Example 7	Example 8
$P(ERR)$	0.666667	0.100000	1.000000	0.950000
$P(COR)$	0.333333	0.900000	0.000000	0.050000
$H(X)$	1.584963	1.584963	1.000000	1.584963
$H(Y)$	1.584963	1.584963	1.000000	1.584963
$H(XY)$	3.169926	2.153959	1.000000	2.153959
$H(X:Y)$	0.000000	1.015967	1.000000	1.015967
RIT	0.000000	0.641004	1.000000	0.641004

Table A3: Summary Statistics: Examples 5-8

	Example 9	Example 10	Example 11	Example 12	Example 13	Example 14
$P(ERR)$	0.015152	0.020202	0.020202	0.020202	0.010204	0.035714
$P(COR)$	0.984848	0.979798	0.979798	0.979798	0.989796	0.964286
$H(X)$	3.459433	3.459433	3.459433	3.459433	3.459090	3.459090
$H(Y)$	3.485449	3.503489	3.485974	3.505207	3.458681	3.543424
$H(XY)$	3.542540	3.560579	3.571994	3.571994	3.515945	3.591750
$H(X:Y)$	3.402342	3.402342	3.373413	3.392646	3.401826	3.410764
RIT	0.983497	0.983497	0.975135	0.980694	0.983445	0.986029

Table A4: Summary Statistics: Examples 9-14