COMPARING THE RECOGNITION PERFORMANCE OF CSRs: IN SEARCH OF AN ADEQUATE METRIC AND STATISTICAL SIGNIFICANCE TEST

Helmer Strik, Catia Cucchiarini, Judith M. Kessens

A^{2}RT, Dept. of Language & Speech, University of Nijmegen, The Netherlands

{Strik, C.Cucchiarini, J.Kessens}@let.kun.nl; http://lands.let.kun.nl/

ABSTRACT

In this paper a new measure of recognition accuracy is introduced which can be used when comparing the performance of two speech recognizers, to establish which is the better one. This metric combines the advantages of previous measures, but excludes their disadvantages. Essentially, the metric is an attempt to quantify the degree of recognition accuracy for each sentence, thus obtaining a more informative measure than either correct or incorrect, in such a way that the statistical significance of the observed differences can be tested. The advantages of our assessment method are illustrated on the basis of both artificial and real performance data of different recognizers.

1. INTRODUCTION

In speech recognition research it is often necessary to compare the performance of two continuous speech recognizers (CSRs), or two versions of the same CSR, to establish which is the better one. This kind of evaluation is known as performance evaluation [4] and is characterized by three elements: a criterion, a measure and a method. In addition to performance evaluation, adequacy evaluation and diagnostic evaluation can also be distinguished [4], but these will not be addressed in this paper.

In automatic speech recognition (ASR) the criterion is recognition accuracy. The measure, in general, is word error rate (WER) and the method consists in establishing agreement at word level between the string of recognized words and what was actually spoken by using a dynamic programming (DP) algorithm. Applying statistical significance tests to such performance measures is not straightforward, as will be discussed in the current article. This is probably one of the reasons why statistical significance is rarely reported in ASR research. However, some measure of statistical significance would be welcome, because otherwise it is unclear how much importance should be attached to the differences in performance observed between the various recognizers. Moreover, this would make it easier to compare (the significance of) the results of your own research with those of others, which usually is quite difficult [6].

Occasionally, a different measure of recognition accuracy is used such as sentence error rate (SER), because it is considered to be more relevant for the task in question, as in the case of credit card number recognition, but maybe also because this measure is more amenable to statistical significance testing. The problem with SER is, though, that it is a more global measure than WER and is therefore less informative about the actual differences in performance between two recognizers.

The ultimate goal of our research is to stimulate the use of statistical significance tests when comparing the performance of CSRs. The main goal of the present paper is to introduce a new metric for measuring recognition accuracy which is suitable for determining statistical significance.

This paper is organized as follows. In section 2 we address evaluation of recognition accuracy at word level. For the sake of clarity we discuss two distinct measures that are often confused with each other. In section 3 we deal with recognition accuracy at sentence level. Again two different measures are considered. In section 4 an alternative measure of recognition accuracy is introduced. Subsequently, in section 5, the pros and cons of all measures of recognition accuracy presented in this paper are discussed and illustrated on the basis of fictitious and real data. Finally, in section 6 we draw some conclusions and make suggestions for future research.

2. EVALUATION AT WORD LEVEL

Let us start by describing the usual procedure used to obtain WER. Each CSR is first trained and then tested with a test corpus. The output of the test corpus is a string of recognized words (RECOG). The latter string is compared to what was actually spoken (SPOKEN). This is done by means of a DP alignment of SPOKEN with RECOG at the word level. The DP alignment reveals the differences between the two strings: substitutions (sub), deletions (del) and insertions (ins). Then WER can be calculated:

WER = (#sub + #del + #ins) / Nw, Nw = # words.

In order to quantify the difference between the two CSRs several measures have been used, amongst others the following two measures [6]:

- absolute difference: Δabs = WER1 – WER2
- relative difference: Δrel = (WER1 – WER2)/WER1

As is clear from the above formulas, WER is a global measure of accuracy that expresses the proportion of words that have been recognized correctly. However, this is not the only measure of accuracy that can be computed at word level. On the basis of the DP alignment of SPOKEN with RECOG it is also possible to express whether each of the words in question was recognized correctly or not. This calculation yields another measure that we will call WCI: Word Correct or Incorrect. For each word a 1 is scored if the word was misrecognized and a 0 if the word was correctly recognized. This would yield a list of Nw numbers for each CSR.

WCI is not introduced here because we consider it to be a valid alternative to WER, but rather to make it clear that this is a different measure with different properties from the point of view of statistical significance testing: WER is just one number, while WCI is a list of Nw numbers. As a matter of fact, the two lists obtained for the two CSRs to be compared could subsequently be used as input to statistical tests (such as McNemar), to determine the differences between the two CSRs and their statistical significance. There are two problems with WCI, though:

1. The application of statistical tests (such as McNemar) requires that the observations be independent. However, in the case of WCI the observations cannot be assumed to be independent for various reasons:

- during decoding an optimization at sentence level is carried out, in which, generally, a language model is used;
- a DP algorithm is used to align SPOKEN with RECOG;
- cross-word processes exist (like, e.g., coarticulation).

2. Insertions are problematic because it is not clear to which word they should be assigned.

To circumvent the latter problem, insertions are sometimes omitted from the evaluation, but this is obviously not a real solution because insertions are important and should be taken into account during evaluation. Given these two problems, we have to conclude that there probably are no suitable statistical significance tests for WCI. Since the two problems mentioned above do not exist at sentence level, researchers have been using evaluation metrics at sentence level, which can be statistically tested for significance, as will be explained in the following section.

3. EVALUATION AT SENTENCE LEVEL

In the previous section we have seen that at word level the measures WER and WCI can be calculated. Similarly, at sentence level we have ‘Sentence Error Rate’ (SER) and ‘Sentence Correct or Incorrect’ (SCI). Like WCI, SCI can have two values, 0 or 1:

- 0 when there are 0 errors; i.e. the sentence has been recognized completely correctly.
- 1 when there are 1 or more errors.

The DP alignment of SPOKEN with RECOG (i.e. the alignment at the word level) can be used to determine for every utterance whether it has been recognized completely correctly or not. This will yield Ns values (Ns = total number of sentences). In contrast with WCI, the zeros and ones of SCI are independent, and there are no insertions. SCI has been used in combination with the McNemar test [2, 3, 5]. Calculating SER on the basis of SCI is straightforward:

SER = sum(SCI)/Ns.

Since SCI is amenable to statistical significance testing, it seems that it should be preferred to WCI. However, SCI has some other disadvantages which will be discussed in section 5.

4. A NEW METRIC

The metric we propose is called ‘Number of Errors per Sentence’ (NES). Given the DP alignment of SPOKEN with RECOG, the value of NES for every sentence can be obtained by summing the number of substitutions, deletions and insertions in that sentence:

NES = #sub + #del + #ins (per sentence).

Like SCI, NES contains Ns independent values, and no insertions. Given that the observations are paired, NES can be used in combination with several paired statistical significance tests which are suitable for variables from the nominal through ordinal and up to the interval measurement level, such as the Signed Pair test, the Wilcoxon Signed Rank test, and the T test.

In the rest of this paper we will focus on the metric and we will use it in combination with one appropriate statistical significance test, i.e. the Wilcoxon Signed Rank (WSR) test.

5. COMPARISON OF METRICS

The obvious question here is: Why introduce NES? This question can best be answered by comparing NES to the other four metrics mentioned above: WER, WCI, SER, and SCI)

WER is the metric which has been used most often. The most important reason for this is probably that WER is a global measure, i.e. with one number one can describe the performance of a CSR (for a certain test set). This is also the case for SER, but, in general, one is more interested in the percentage of words that have been recognized correctly than in the percentage of sentences correctly recognized. If one assumes that the distribution is normal, then WER and Nw can be used to calculate confidence intervals. In turn these confidence intervals have been used to calculate the statistical significance of the differences between the WERs of two CSRs [see e.g. 1, 3].

Although conciseness is an advantage of WER, this metric also has some disadvantages. The global number WER does not reveal much about the recognized strings of words on which it is based. Very different recognition results can yield similar WER values [7]. In turn, these similar WERs result in similar confidence intervals. When two of these confidence intervals are compared, in order to test whether two WERs are significantly different, the underlying distribution of the errors is not taken into account. Furthermore, these tests with confidence intervals are not very powerful (see, e.g., Table 2b). Therefore, this method (WER in combination with confidence intervals) does not seem to be optimal for determining statistical significance. SER has similar advantages and disadvantages, but – as mentioned above – is used much less often than WER.

WCI is more detailed than WER. However, given the properties mentioned in section 2, there probably are no suitable statistical significance tests for WCI. SCI and NES, on the other hand, can be used in combination with several statistical significance tests. Therefore, the real comparison is that between SCI and NES. For this comparison we will use both artificial examples and some real data (i.e. from CSRs of our own research). We will start with the artificial examples in Table 1.

First of all, it should be noted that SCI is not detailed enough and is therefore little informative about the differences in performance between two CSRs. While a zero value means that the sentence in question does not contain any error, a value of one can mean a lot of different things, varying from one error through all gradations up to all words recognized incorrectly. This clearly appears from the examples presented in Table 1.

utterance	**1**	**2**	**3**	**4**
CSR1				
#sub	1	2	3	1
#del	1	2	3	0
#ins	1	2	3	0
NES	3	6	9	1
SCI	1	1	1	1
CSR2				
#sub	1	1	1	0
#del	0	0	0	1
#ins	0	0	0	0
NES	1	1	1	1
SCI	1	1	1	1
CSR1-CSR2				
ΔNES	2	5	8	0
ΔSCI	0	0	0	0

Table 1: A comparison of artificial recognition results.

NES, on the other hand, is in line with our intuitions that in a sentence containing 2 errors recognition accuracy is higher than in a sentence containing 7 errors, whereas this difference would not be revealed by SCI. Of course, one could argue that not all errors are equally serious and that, therefore, just counting errors without making further distinctions is not satisfactory. This may be true, but this observation does not alter the fact that, on the basis of our intuitions, NES does constitute an improvement with respect to SCI (and the other metrics). The possibility of making distinctions among error types will be further discussed in section 6.

Let us now compare CSR1 with CSR2 for the examples given in Table 1. For utterance 1 we observe an improvement from 3 to 1 recognition errors. Since errors are present for both CSRs, SCI is 1 for both utterances, and thus ΔSCI is 0. NES does decrease from 3 for CSR1 to 1 for CSR2, and ΔNES is 2. In other words, the differences in recognition results for utterance 1 are reflected in ΔNES but not in ΔSCI. Analogously, the changes for utterances 2 and 3 are 'noticed' by ΔNES, but not by ΔSCI. Not only does ΔNES reflect the differences for the utterances 1 to 3, ΔNES also reveals that the magnitude of the improvements increases when going from utterance 1 to 3.

These examples show that ΔNES does measure some differences which ΔSCI does not measure. However, it should be noted that there are some changes that are not even reflected in ΔNES, e.g. like those for utterance 4. As a matter of fact, we can see that NES makes no distinction between substitutions, deletions and insertions. However, this distinction is not made by SCI either and, consequently, the changes in utterance 4 are not reflected in ΔSCI either.

Let us now look at some real data, i.e. compare the output of two CSRs. In Table 2a we first present some descriptive statistics, i.e. total number of sentences and words in the test set, SER and WER for the two CSRs, and the two different ways to express the difference in the WERs that were mentioned above.

2a: Descriptive statistics.	
Ns	5000
Nw	16357
SER1	26.54%
SER2	25.92%
WER1	15.64%
WER2	14.67%
Δabs	0.98%
Δrel	6.25%

2b: Combinations of metrics and statistical tests.		
metric	statistical test	p
WER	confidence intervals	5.8%
SCI	McNemar	11.3%
SCI	WSR	10.2%
NES	WSR	0.9%

Table 2: A comparison of two CSRs.

The level of significance of the difference is calculated by using several metrics in combination with statistical tests. The results are presented in Table 2b. For WER in combination with the confidence intervals we used the formula presented in [1]. It can be observed that for none of the first three combinations the difference is significant at the 5% level. NES in combination with WSR, on the other hand, does indicate that the difference is highly significant ($p < 1\%$). The differences between the results for NES and SCI can be explained by the fact that for the selected combination of CSRs, there are a lot of utterances for which the number of errors is reduced but does not become zero. Consequently, for these utterances SCI remains one, while NES is reduced (as was the case for utterances 1 to 3 in Table 1). SCI finds 195 improvements and 164 deteriorations, while NES finds many more: 345 improvements and 289 deteriorations. This is also reflected in the (relatively) small reduction in SER (SER = sum(SCI)/Ns) and the larger relative reduction in WER (WER = sum(NES)/Nw).

6. FINAL REMARKS

In this paper we have discussed several metrics that can be used to express recognition accuracy when comparing the performance of two CSRs. In particular, we have discussed these metrics with respect to their informative properties and their possibilities of being submitted to statistical significance tests. It turned out that many of these metrics are not really satisfactory, either because they are not informative enough, or because there is no suitable significance test. More precisely, for determining the significance of the differences between two CSRs, the metrics WER, SER, and WCI are less suitable than SCI and NES. SCI and NES can be used in combination with several statistical significance tests, like the Signed Pair test, the Wilcoxon Signed Rank test, and the T test. We have showed that NES has many advantages compared to SCI, and also to the other three metrics (WER, WCI, and SER). NES is more in line with our intuitions about differences in recognition accuracy and is able to detect differences that other metrics cannot detect, thus providing more information about the differences in performance between two CSRs. In addition, NES can be used in combination with statistical tests that are more powerful than the McNemar test and is therefore able to detect more significant differences. In other words, NES seems to be more suitable for comparing two CSRs than the other metrics discussed here.

A question that remains to be answered is which statistical significance test should be used in combination with NES. Possible tests are the Signed Pair test, the Wilcoxon Signed Rank test, and the T test, as mentioned above. A short word about these tests is in order. The Signed Pair is a nonparametric test that only looks at the direction of the change (the sign). WSR also is a nonparametric test, but the difference with the Signed Pair test is that in WSR the ranking is done on the basis of the direction and the magnitude of the change. Finally, the T test is a parametric test that also takes the magnitude of the difference into account. The T test and WSR are more powerful than the Signed Pair test and the McNemar test. WSR has about 95% of the power of the T test, if all the assumptions of the T test are met. Since this is probably not the case here, we have decided to use WSR in the current article. However, further research should reveal whether WSR is really the optimal test for NES.

At this point, we would like to go back to the point concerning error seriousness which was mentioned in section 5. As a possible drawback of NES one could mention that it makes no distinction among error types. It is important to note, though, that this kind of information could be added to NES, thus obtaining a more informative metric. First of all, if one were convinced that substitutions, deletions and insertions have a different impact on recognition accuracy and would like to express this in a measure of recognition accuracy, one could compute separate measures for the three error types. This gain in informativeness would be accompanied by a loss in conciseness, with consequent extra problems for determining statistical significance. Another possibility would be to assign different weights to the three types of errors:

NES = Ws*#sub + Wd*#del + Wi*#ins (per sentence).

Currently, Ws = Wd = Wi = 1, but different values could be used. The problem in this case would be to find agreement on the relative seriousness of the three error types, which, probably, cannot be established in absolute terms, because it differs from case to case (e.g. depending on the application, the task the CSR is used for).

To summarize, more informative measures of recognition accuracy can be devised, if necessary, but these may introduce new problems from the point of view of evaluation and statistical testing. Therefore, for the time being, we decide to stick to NES which constitutes a considerable improvement in informativeness, without presenting insurmountable problems for statistical testing.

ACKNOWLEDGEMENTS

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

REFERENCES

1. Ferreiros, J., and Pardo, J.M. “Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations”, *Speech Communication,* Vol. 29: 65-76, 1999.

2. Gillick, L., and Cox, S.J. “Some statistical issues in the comparison of speech recognition algorithms”, *Proc. ICASSP:* 532-535, Glasgow, May 1989.

3. Harborg, E., *Hidden Markov Models Applied to Automatic Speech Recognition,* Ph.D. thesis, Norwegian Institute of Technology, Trondheim, 1990.

4. Hirschman, L., and Thompson, H.S. “Overview of evaluation in speech and natural language processing”, In: R. Cole et al. (eds.) “Survey of the State of the Art in Human Language Technology”, Cambridge University Press, 1997.

5. Pallett, D., Fiscus, J., Fisher, W., and Garofolo, J. “Benchmark tests for the DARPA spoken language program”, *Proc. of the 1993 ARPA workshop*: 7-18, 1993.

6. Strik, H., and Cucchiarini, C. “Modeling pronunciation variation for ASR: A survey of the literature”, *Speech Communication,* Vol. 29: 225-246, 1999.

7. Wester, M., Kessens, J.M., and Strik, H. “Pronunciation Variation in ASR: Which Variation to model?”, *Proc. ICSLP-2000,* Beijing, 2000.