WITHIN-SPEAKER VARIABILITY OF THE WORD ERROR RATE FOR A CONTINUOUS SPEECH RECOGNITION SYSTEM

David A. van Leeuwen and Herman J. M. Steeneken
Electronic mail: {vanLeeuwen, Steeneken}@tm.tno.nl
TNO Human Factors Research Institute.
Postbus 23,
3769 ZG Soesterberg,
The Netherlands.

ABSTRACT

The variance of the performance of a continuous speech recognition system subjected to replica utterances of the same sentence spoken by the same speaker has been investigated. In an experiment with three different speech recognition systems in three different languages with two different grammar conditions it is shown that the sentence word error rate has a variance that can be described in terms of binomial statistics. The distribution of the measured variance shows a remarkable correspondence to the parameter-free theoretical distribution. It is therefore concluded that for the word error rate of a *continuous* speech recognition system binomial statistics apply.

INTRODUCTION

The word error rate (sometimes expressed in its complement, the accuracy) is the most widely used measure of the performance of speech recognition systems. Traditionally, for isolated word recognizers this measure has been one which leaves little argument for interpretation, but for continuous speech recognition systems the situation is more complex. Because of the nature of natural speech the words are connected to a long string. This makes it somewhat difficult to pinpoint the exact location of an error in case of misrecognition and consequently makes it hard to count the number of erroneous words. Evaluating the correctness of utterance as a whole, measured in the utterance (or sentence) error rate resolves this problem. However, this measure needs much more speech material before an accurate figure is found, and researchers often use the word error rate because it is more sensitive to small changes in the performance of the speech recognition system.

One of the questions we want to address in this paper, is how accurate a measurement of the word error rate is for a continuous speech recognition system. For a representative evaluation, one generally wants to have a wide coverage of language, and in case of a speaker independent system, a wide coverage of speakers. Because both sets are virtually infinite in size, for each evaluation new samples are drawn from the sets of language material (sentences) and speakers. If there are ways to quantify the accuracy of a word error rate measure, and objective ways to calibrate the 'difficulty' of the test material [1], a new evaluation can successfully be compared to an earlier one.

EXPERIMENTAL SETUP

In order to study the inherent variability of the performance of a continuous speech recognizer, we performed a test with no variability in speaker and spoken text. This experiment was carried out as an additional test in the project SQALE, which was a project that compared speech recognition in different European languages and for different systems [1, 2]. The variability in speaker and speech content was made zero by having a speaker read out the same sentence several times, of which we call the individual utterances *replicas* of the same sentence. (These replicas can in principle be used to measure the within-speaker variability.) The replicas were recorded during a recording session of the evaluation test of SQALE, and were spread among the normal evaluation sentences. The speakers were prepared for the occurrence of replicas, and were requested to read out a replica as if it was the first occurrence in order to make the utterances as much alike as possible. We chose for 5 replicas of one sentence for each recorded speaker; more replicas might have stretched the subject's acceptance limits too far, and we did not not want that the reading style of the other (evaluation test) utterances was influenced by this test.

Table. The number of sentences available, for each language. Each speaker, having its own sentence, uttered 5 replicas. The number of speech recognition systems available per language is also indicated, as well as the amount of measurement points resulting.

Language	American English	British English	German
sentences	3	7	10
systems	3	3	2
grammars	2	2	2
data points	18	42	40

The replica utterances were recorded in three different languages, in amounts according to the ta-

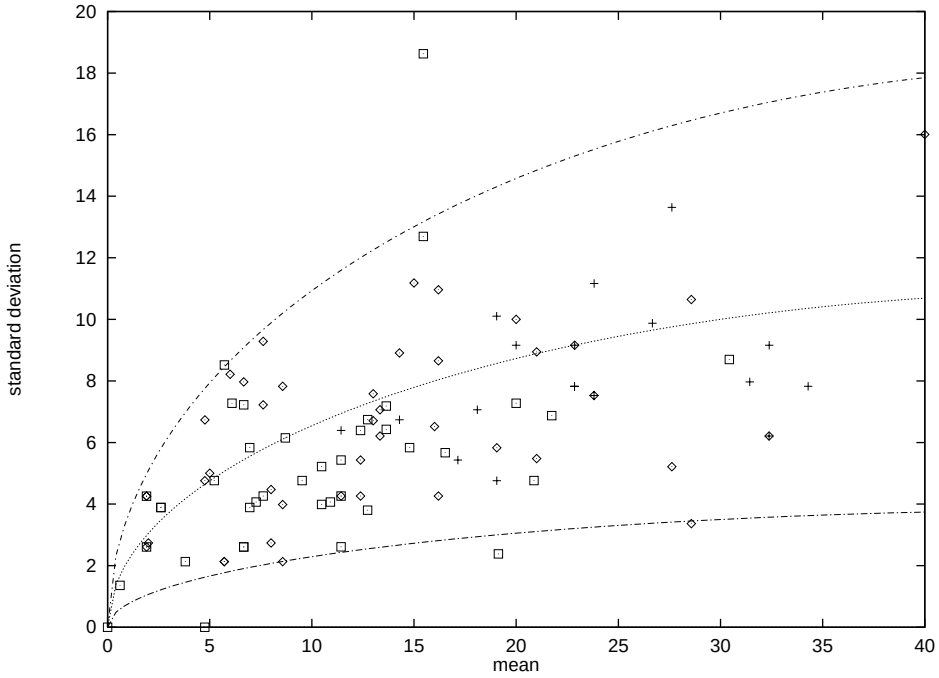

Figure 1. All data of the replica experiment. For 5 identical sentences uttered by the same speaker, the standard deviation of the word error rate is plotted versus the mean word error rate. The data points comprise two grammar conditions, three systems and three languages. The expected relation for a binomial distribution (central line), along with the 2.5 % and 97.5 % percentile cut-offs are plotted as well. ⊡ German, ⋄ British English, + American English.

ble. The number of available speakers is not constant across languages, which is due to limited availability of resources within the project. Three partners in SQALE joined in the experiment, in several languages, using both a bigram and a trigram language model. Due to the limited and incomplete 'design' of the experiment, as shown in the table, we will consider all differences in language, system, grammar and speaker (hence sentence) merely as parameters that widen the range of values the true word error rate of a sentence can have.

RESULTS

All utterances were submitted to the recognition partners, who ran their various recognition systems on the material and sent back the recognition results to us, having the function of independent coordinator within the project. We performed standard alignment and scoring of the recognition results, and determined for each sentence the mean word error rate $\hat{w}$ and the standard deviation $\hat{s}_w$ (the square root of the variance) from the word error rates w_i found for the $N = 5$ replicas. We used the usual estimators

$$\hat{w} = \frac{\sum_{i=1}^{N} w_i}{N}; \qquad \hat{s}_w^2 = \frac{\sum_{i=1}^{N} (w_i - \hat{w})^2}{N-1}. \qquad (1)$$

In figure 1 we plotted $\hat{s}_w$ as a function of $\hat{w}$ for each of the 100 data points. At a first glance the data points seem to be scattered rather wildly around the theoretical dependence

$$\sigma_{\mathrm{ex}}(w) = \sqrt{\frac{w(1-w)}{N_w}} \qquad (2)$$

that expresses the standard deviation $\sigma_{\mathrm{ex}}(w)$ expected given the true word error rate w, the number of words per sentence N_w and assuming a binomial distribution of w. (This assumtion is accepted for isolated speech recognition.) It is legitimate to use one curve to compare all data points, because all sentences used in the test had $N_w = 21$ (except for a few of the German sentences which all have N_w ranging 20–23, but we will ignore the small error made here).

The reason why the data points deviate this much from the expected line is that the estimation of the variance, based on only 5 replicas, is in fact rather inaccurate. In figure 1 we have also indicated the 2.5 % and 97.5 % percentile curves for an $N = 5$ standard deviation determination. In fact, only 4 out of the 100 data points do not fit within this 95 % confidence interval. In order to investigate the validity of the binomial assumption made in eq. (2) we now look how the measured standard deviation is distributed with respect to the expected standard deviation.

Because the standard deviation estimate is a sum over squares of the statistic w, $\hat{s}^2$ is distributed according to a χ^2 distribution with $N-1$ degrees of freedom [3]. In fact, one expects that

$$\chi^2(w) = \frac{N-1}{\sigma_{\mathrm{ex}}^2(w)} \hat{s}^2 \qquad (3)$$

is distributed as a χ^2_{N-1}-distribution. In figure 2 we have plotted a histogram of this $\chi^2(\hat{w})$ according to

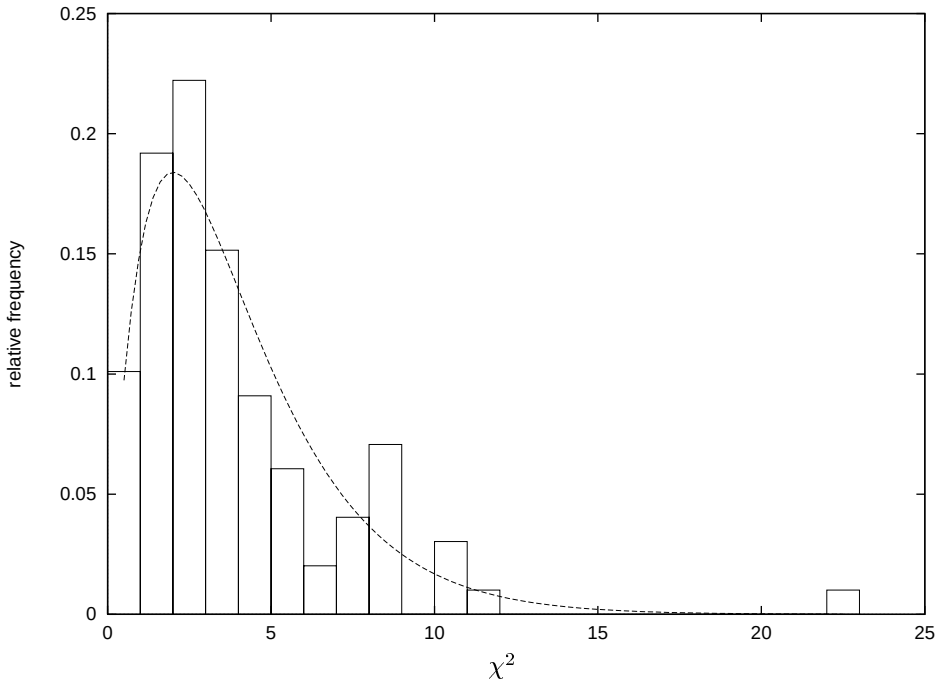

Figure 2. Distribution of the quantity $\chi^2(\hat{w})$ of eq. 3, which is calculated from the data points plotted in figure 1. The curve drawn is the theoretical distribution $f(\chi^2; 4) = \frac{1}{4}\chi^2 e^{-\chi^2/2}$.

eq. 3 along with the theoretical distribution χ^2_4. The correspondence is striking, especially considering the fact that the theoretical curve *contains no fitting parameters*. We are therefore led to believe that the variance of the word error rate w that is left when a continuous speech recognition system is subjected to the same sentence uttered by the same speaker several times can be described entirely by the variance of a binomial statistic w.

DISCUSSION

It is surprising that replica utterances, that are perceptually identical, show such a large variance in recognition rate as indicated in figure 1. We can think of two mechanisms that influence this variance. On the one hand, one could expect that if a recognizer makes systematic errors (*e.g.*, recognizing 'a' for 'the') it would have a relative low variance in the error rate. On the other hand, we can expect that genuine within-speaker variance would be an additional source of variability in word error rate. Analyzing the variance, we were not able to identify any of the observed variance to be caused by within-speaker variance, which was our original goal in doing this experiment. We cannot exclude that the two effects cancel, but the evidence found in both figures suggest that the variability of replica sentences is due to the binomial nature of the word error rate.

This experiment is performed using *continuous* speech recognition systems, which gives rise to an even more striking conclusion: the whole process of continuous speech production, recognition and evaluation contain inter-word influences such as coarticulation effects, grammar and arbitrary alignment decisions, and yet, the word error rate can be considered a binomially distributed statistic as if each word has been produced and recognized independently of the others. This has an important consequence:

> From the average word error rate w and the number of words in the evaluation test set N, an inherent inaccuracy e of the measurement for w can be determined, according to
>
> $$e = \sqrt{\frac{w(1-w)}{N}}. \qquad (4)$$

In comparing two different speech recognition systems, special significance test exist [4, 5, 6], that take advantage of errors common to both systems. These allow to find differences smaller than e expressed above to be significant. If, however, the test set is not common to both evaluations, an error measure according to eq. 4 gives a minimum estimate of the inaccuracy of the evaluation. Because the error measure can become relatively large for a small test set, as is the case in the comparison between different speakers having different sets of test sentences, we want to warn speech recognition researchers not to put too much weight on differences between the results of two different systems or speakers smaller than found by eq. (4).

We also want to argue that when speech recognition error rate results are published, apart from the usual experimental conditions, the inherent inaccuracy of the measurement (as expressed in eq. 4) should be mentioned, even if it concerns a continuous

speech recognition system. In presenting the results of SQALE [1], the only systems that differed significantly by the tests described above [5], also had their word error rates differing more than their mutually summed inaccuracy figures.

We have learnt from this experiment that if one wants to measure a genuine effect of within-speaker variability on the word error rate by using replica utterances, the number of replicas must be relatively large in order to narrow down the inaccuracy range shown in figure 1.

ACKNOWLEDGEMENTS

We want to thank the partners in the SQALE project that kindly ran the replica utterances: LIMSI-CNRS and Cambridge University Engineering Department's HTK and Abbot. Part of this research has been sponsored by the CEC, under contract number LRE 62-058.

REFERENCES

1 Herman J. M. Steeneken and David A. van Leeuwen. Multi-lingual assessment of speaker independent large vocabulary speech recognition systems: The SQALE project. In *ESCA Proc. Eurospeech*, Madrid, September 1995.

2 S. J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvain, D. J. Kershaw, L. Lamel, D. A. van Leeuwen, D. Pye, A. J. Robinson, H. J. M. Steeneken, and P. C. Woodland. Mutilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language*, 11:73–89, 1997.

3 William L. Hays. *Statistics*. Holt, Rinehart and Winston, Inc., 1963.

4 L. Gillick and S. J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *IEEE Proc. ICASSP*, pages 532–535, Glasgow, 1989.

5 Alvin Martin et al. Statistical significance tests for speech recognition benchmark tests. Technical report, National Institute of Standards and Technology (NIST), July 1995. Draft version, kindly provided by Dr. Martin.

6 Jeffrey N. Marcus. Significance tests for comparing speech recognizer performance using small test sets. In *ESCA Proc. Eurospeech*, pages 465–468, Paris, 1989.