

INSIGHTS FROM THE BROADCAST NEWS BENCHMARK TESTS

Walter Liggett and William M. Fisher

National Institute of Standards and Technology
Information Technology Laboratory
Building 225, Room A216
Gaithersburg, MD 20899
E-mail: walter.liggett@nist.gov

ABSTRACT

The broadcast news benchmark tests have potential as a source of ideas for improving continuous speech recognition systems. This paper presents a data analysis method for uncovering such ideas and applies the method to the 1996 and 1997 DARPA CSR Hub-4 results. The method is based on a latent variables model instead of a more familiar regression model. The method identifies certain portions of the test material that result in wide performance differences among systems. Such portions, because some systems could handle them and others could not, are worth thinking about in terms of what system features lead to the performance differences. Identification of specific system differences that are responsible for performance differences may lead to system improvements.

1. INTRODUCTION

Benchmark tests of continuous speech recognition systems usually entail having each system transcribe the same selection of speech. In the case considered here, the selection is from broadcast news. The system transcriptions are compared to an assumed perfect transcription and system transcription errors identified. System-to-system comparisons of these errors can be used to choose the best system from among those tested. In addition, such comparisons can be used to gain insights into the effects of system differences on performance. Such insights are the goal of the method presented in this paper.

The program of data interpretation we have in mind involves description of differences in system transcription errors, description of differences in system features, and finally development of relations between these two descriptions. We focus on a method for error description because our knowledge of system features is insufficient. Hopefully, those who have in depth system knowledge will be able to use either the error description given here or an error description from the method given here to obtain system insights.

Fundamental to the development of speech recognition systems is comparison of alternatives in terms of observed word error rates. What alternatives should a developer compare? This paper provides help in answering this

question, help in finding alternatives that lead to major system improvements. This help can come in the form of segments of speech or speakers to which a developer can listen. This help can also come in the form of segment categories that a developer can study. The segments, speakers, or segment categories provided are the most important for a developer to consider because they are portions of speech that some systems transcribe better than others. Thus, these portions of speech point to improvement that can be made in at least some systems. A developer that pays attention to these portions avoids wasting time on portions of speech that any system can easily transcribe or that no system can transcribe well.

The analysis method presented here is not based on regression, that is, modeling system performance in terms of (potentially) manifest variables such as signal-to-noise ratio, speaker gender, or out-of-vocabulary words [1]. Rather, the analysis method is based on latent variables, which are not given as values for each segment in the data set [2]. Generally, the method consists of looking for latent variables that have a large effect on performance differences among systems. After such variables have been identified, they can be characterized in ways helpful in pinpointing subsystems in need of improvement.

We begin the analysis with system word error rates for partitioned segments. Segments are turns in speaking by a single speaker, and partitioned segments have been further divided at points where the speech condition, the so-called focus condition, changes. The word error rates and segment characteristics were derived through adjudicating differences among three transcribers and annotators. Think of the word error rates for systems and segments as a two-way table with rows corresponding to systems and columns to segments. The method presented analyzes this table, that is, decomposes the table into components that when added back together give the original word error rates. The decomposition involves a component for overall system performance and a component for segment difficulty. Further, there is a component that reflects the degree to which segment error rates are closely or not so closely aligned with overall system performance. Beyond the overall system performance, there may be other

differences in system performance that appear when a subset of the segments is considered but are offset by other segments in the overall performance. Such differences can be thought of as the result of latent variables. Finally, what remains is what looks like random variation. The components in the decomposition provide a starting point for an effort to build relations between system features and performance. Such an effort may be time consuming. Thus, in performing this decomposition, there is a need to identify those components distinct enough to be worthy of careful examination.

There are various tables to which our analysis can be applied. Section 2 contains the analysis of the system by segment tables for 1997 and 1996. Instead of segments, we can form tables of speaker word error rates or focus condition word error rates and apply our analysis. As discussed in Sections 3 and 4, collapse of the columns to speakers or focus conditions can make some components more distinct and thus more clearly worthy of consideration. In Section 5, we discuss a somewhat different topic, inference from the 1997 results. We discuss the question of how the systems would compare if they were used to transcribe a much larger collection of speech.

2. SYSTEM BY SEGMENT RESULTS

2.1. 1997 Results and Analysis Method

Consider the 1997 test results for the focus conditions F0 (Baseline Broadcast Speech) and F1 (Spontaneous Broadcast Speech), the two focus conditions agreed on as the emphasis for 1997. The system word error rates for these two focus conditions combined are given in the second column of Table 1. We have omitted the results from OGI because their word error rate for F0 and F1 is 31.6, which is so much different from the other results that the OGI results might dominate the analysis.

Table 1. 1997 Results for F0 and F1 Combined.

System	Word Error Rate (WER)	Centered WER \hat{x}_i
bbn	13.5	-1.6
cmu	17.1	2.0
cu-con	18.7	3.7
cu-htk	11.7	-3.4
dragon	16.8	1.7
ibm	12.7	-2.3
limsi	13.2	-1.9
philips	16.7	1.7
sri	15.2	0.1

The system word error rates in Table 1 can be viewed as

weighted averages of the word error rates for system and segment. Letting the index i range over the m systems and the index j over the N segments, we denote the system-segment word error rates by Y_{ij} . For segment j , the weight is the number of words in the reference transcription of the segment, which we denote by n_j . The second column in Table 1 is given by

$$\bar{Y}_i = \frac{\sum_{j=1}^N n_j Y_{ij}}{\sum_{j=1}^N n_j}$$

Note that, for each system, this equation is the ratio of the total number of errors for all segments divided by the total number of words for all segments. The third column in Table 1, labeled \hat{x}_i , gives system word error rates centered at the average over the systems

$$\hat{x}_i = \bar{Y}_i - \frac{1}{m} \sum_{i=1}^m \bar{Y}_i.$$

In the analysis presented here, these centered word error rates portray the overall ability of each system.

To complete what might be thought of as the simplest possible analysis of the system by segment word error rates, we include a term representing segment difficulty. For simplicity, we estimate this term without centering; we take as the estimate

$$\hat{\alpha}_j = \frac{1}{m} \sum_{i=1}^m Y_{ij}$$

The model

$$Y_{ij} = \alpha_j + x_i + \epsilon_{ij}$$

where ϵ_{ij} is regarded as zero-mean random error with variance that depends on n_j , may adequately represent the word error rates, but, of course, one can think of reasons why it might not.

One extension is addition of a term that describes variation from segment to segment in the effect of the overall system abilities [2]. This extension leads to the model

$$Y_{ij} = \alpha_j + (1 + \beta_j) x_i + \epsilon_{ij}$$

One can think of the β_j as representing the degree to which a segment can distinguish the varying abilities of the systems. The β_j can be estimated by

$$\hat{\beta}_j = \frac{\sum_{i=1}^m (Y_{ij} - \hat{\alpha}_j - \hat{x}_i) \hat{x}_i}{\sum_{i=1}^m \hat{x}_i^2}$$

Note that this estimate is just a linear regression of adjusted segment word error rates on the overall system abilities $\hat{\mathbf{x}}_i$.

Characterization of the segments that have the highest values of $\hat{\beta}_j$ may provide insights into system improvement since these segments most strongly distinguish system performance. However, before attempting such characterization, one should check to see if the variation in the $\hat{\beta}_j$ is large enough to make the effort worthwhile. To do this, we assume that the variance of ϵ_{ij} is proportional to $1/n_j$. This assumption is not unreasonable but cannot be completely justified either. Under the assumption that $\beta_j = 0$, the variance of $n_j^{1/2} \hat{\beta}_j$ is estimated by

$$\frac{1}{(N-1)} \sum_{j=1}^N n_j \hat{\beta}_j^2$$

Whatever the values of β_j , the variance of $n_j^{1/2} \hat{\beta}_j$ is estimated by

$$\frac{\sum_{i=1}^m \sum_{j=1}^N n_j (y_{ij} - \hat{\alpha}_j - (1 + \hat{\beta}_j) \hat{\mathbf{x}}_i)^2}{((m-2)(N-1)) \sum_{i=1}^m \hat{\mathbf{x}}_i^2}$$

We can compare the former to the latter to see if the variation in $\hat{\beta}_j$ is so small that it is not worth characterizing. The ratio of these two estimates is an F statistic with $N-1$ and $(m-2)(N-1)$ degrees of freedom. Thus, the ratio can be compared to a critical point obtained from a table of the F distribution.

One might expect that deviation from the assumption that the variance of ϵ_{ij} is proportional to $1/n_j$ is most severe for short segments. To avoid undue influence by the shortest segments, we compute the F ratio from the 182 segments with $n_j \geq 30$. The total number of F0 and F1 segments N is 318. From the resulting value of 1.43, we conclude that characterization of the $\hat{\beta}_j$ is worthwhile.

Plotting $\hat{\beta}_j$ in various ways shows three possible influences of segment type on the segment-to-segment effect of overall system ability. From a plot of $\hat{\beta}_j$ versus n_j , we conclude that segments with fewer than 10 words have higher $\hat{\beta}_j$ and thus can be thought to distinguish the systems more persuasively than those with more words. From a plot of $\hat{\beta}_j$ versus $\hat{\alpha}_j$, we conclude that more difficult segments distinguish the systems more persuasively. Finally, we see that the segments with focus condition F1 distinguish the systems more persuasively than those with focus condition F0. Short segments, difficult segments, and F1 segments largely occur together. For this reason, deciding which of these influences is dominant seems difficult. These influences might suggest types of speech that can be transcribed more effectively than what is being achieved by

systems other than the best. Thus, these observations might suggest where to seek system improvements.

In some cases, adding another term to the model of Y_{ij} is informative

$$Y_{ij} = \alpha_j + (1 + \beta_j) \mathbf{x}_i + \gamma_j \mathbf{z}_i + \epsilon_{ij}$$

In order to distinguish this new term from the ones considered already, we require that

$$\sum_{i=1}^m \mathbf{z}_i = 0$$

and that the vector \mathbf{z}_i be orthogonal to \mathbf{x}_i

$$\sum_{i=1}^m \mathbf{z}_i \mathbf{x}_i = 0$$

The interpretation of this new term has an analogy in intelligence testing where one first considers general intelligence and then verbal ability versus mathematical ability. Similarly, this new term portrays some segments that some systems do well with and other segments that other systems do well with.

We estimate the new term by applying the singular value decomposition to the matrix with elements

$$n_j^{1/2} (y_{ij} - \hat{\alpha}_j - (1 + \hat{\beta}_j) \hat{\mathbf{x}}_i)$$

The rank of this matrix is the smaller of $m-2$ and $N-1$. The singular value decomposition represents this matrix as

$$\sum_k u_{ik} d_k v_{jk}$$

where the columns of the matrix u_{ik} and columns of the matrix v_{jk} are orthonormal and the d_k are non-negative and in descending order. The number of d_k that are non-zero equals the rank of the matrix. The new term, which is obtained from the first one in the sum, is estimated by $n_j^{-1/2} u_{i1} d_1 v_{j1}$. So that the magnitude of $\hat{\mathbf{z}}_i$ is comparable to that of $\hat{\mathbf{x}}_i$, we let

$$\hat{\mathbf{z}}_i = \frac{d_1 u_{i1}}{\left(\sum_{j=1}^N n_j\right)^{1/2}}$$

and

$$\hat{\gamma}_j = n_j^{-1/2} v_{j1} \left(\sum_{j=1}^N n_j\right)^{1/2}$$

For the 1997 F0-F1 data that we have considered so far in this paper, the non-zero values of d_k are 9.3, 8.7, 8.2, 7.9, 7.2, 6.1, and 5.9. Since the first of these values does not stand out from the rest, we conclude that for these data, the

new term is not worth considering. If the first two of these values stood out from the rest, then we could have considered two new terms, but this is not true either. The importance of this new term is clearer in the 1996 F0-F1 data.

2.2. 1996 Results

Consider now the 1996 test results for the focus conditions F0 and F1, results which were discussed last year by Pallett, et al. [3]. The scoring we consider is the one used last year, not the somewhat revised scoring used for the 1997 results. The system word error rates for these two focus conditions, the values of \hat{x}_i , and other values to be discussed below are given in Table 2.

Table 2. 1996 Results for F0 and F1 Combined.

System	WER	Centered WER \hat{x}_i	Contrast \hat{z}_i	Segment by Dole
bbn	30.2	-1.0	-1.4	22.5
cmu	33.8	2.5	7.4	60.5
cu-con	34.3	3.1	-2.8	33.2
cu-htk	27.4	-3.9	1.0	17.8
ibm	30.7	-0.6	-1.7	22.1
limsi	28.0	-3.2	0.1	15.8
sri	34.4	3.1	-2.5	24.5

As above, we first consider $\hat{\beta}_j$, the ability of a segment to distinguish systems. We compute the F ratio using segments with 30 or more words and obtain the value 1.02. From this we conclude that further examination of the $\hat{\beta}_j$ is not worthwhile.

The non-zero values of \hat{d}_k from the singular value decomposition are 9.9, 8.7, 7.9, 7.1, and 6.6. The first of these values stands out from the rest weakly but better than in the case of the 1997 data. What makes the $\sum_j z_i$ term worth considering are the values of \hat{z}_i which are shown in Table 2. Clearly, this term represents some difference between the performance of the CMU system and the others. To pursue this further, we obtain the largest value of $n_j^{1/2} \hat{y}_j$, which points to the segment most responsible for this term. The largest value, for which \hat{y}_j is 3.2, corresponds to a 253 word segment spoken by Senator Dole. The word error rates for this segment are shown in Table 2. Why the CMU system did so poorly with this segment may be worth considering. Thus, this term in our analysis points to a clue that may provide insight into improving system performance.

2.3. Speakers Common to 1996 and 1997

Tables 1 and 2 show that the average word error rate for 1997 is 15.1 and for 1996 is 27.1. Thus, from 1996 to 1997, there

is a 44 percent (relative) decrease in the word error rate. One might wonder how much of this decrease is due to easier material and how much to system improvement. One might conjecture that easier material would correspond to easier speakers and therefore that one should consider speakers that appear in both years. Table 3 shows five speakers, David Brancaccio, Donna Kelly, Leon Harris, President Clinton, and Senator Dole, that are included in both test sets. Along with each speaker is the median of the relative decrease in system word error rate, the median over the seven systems that participated in both years.

Table 3. Speakers Common to Both Years

Speaker	Median Relative Decrease
Brancaccio	37%
Kelly	54%
Harris	48%
Clinton	5%
Dole	9%

We see that the first three speakers show a relative decrease in word error rate comparable to the decrease in the average for all speakers and systems. Assuming that the sites did not anticipate these particular speakers, the results for these speakers suggest that the 1997 decrease in word error rate is largely due to system improvement. The last two speakers do not show a relative decrease as large. Why did the systems improve less for these last two speakers? One possible explanation is the change in source for these speakers from 1996 to 1997. In 1996, all the material was from news broadcasts. In 1997, the contribution of these speakers was a portion of the CSPAN archive of the presidential debates. Further investigation might show this change in source to be responsible for the apparent discrepancy between the five speakers. On the other hand, it may be true that the separate effects of easier material and system improvement on the word error rate cannot be resolved.

3. SYSTEM BY SPEAKER RESULTS

It is well known that performance varies with speaker and that for a particular speaker, performance differs between speech that is previously prepared material being read and speech that is spontaneously formed. In addition, there are other variations in speech that can cause performance for a particular speaker to change. These include rate of speech, grammatical complexity of the material, and use of out-of-vocabulary words. For broadcast news, within a focus condition, it seems plausible that speaker-to-speaker variation is generally much larger than other segment-to-segment variation. For this reason, one might consider applying the foregoing analysis to tables of word error rates by system and speaker.

The word error rates for a particular speaker are obtained from the word error rates for segments spoken by that

speaker. The speaker word error rate can be viewed as a weighted average of the segment word error rates with weighting given by the number of reference words in the segment. This gives the word error rate for a system-speaker category as the number of errors for the category divided by the number of reference words in the category. For this reason, parts of the analysis do not change as one goes from segments to speakers. The system word error rates and the \hat{x}_i do not change. The values of $\hat{\alpha}_j$ and $\hat{\beta}_j$ for the case of speakers are just weighted averages of the corresponding values for the case of segments. Only the part of the analysis based on the singular value decomposition changes in a substantial way. It is possible for \hat{d}_1 to stand out more clearly from the rest when a different grouping such as speaker is used.

We analyze the 1997 system by speaker word error rates separately for focus conditions F0 and F1. The results for F0 are given in Table 4.

Table 4. 1997 Results for F0.

System	Word Error Rate	Centered WER \hat{x}_i	Contrast \hat{z}_i
bbn	11.4	-1.2	-1.0
cmu	14.4	1.8	2.2
cu-con	15.5	2.8	-0.5
cu-htk	9.9	-2.8	-1.2
dragon	13.9	1.2	0.2
ibm	10.3	-2.4	2.8
limsi	11.6	-1.1	-0.7
philips	14.4	1.8	-1.0
sri	12.5	-0.1	-0.7

We see that the system word error rates are not much different from those shown in Table 1 for F0 and F1 combined. There are 49 speakers. The F ratio for the values of $\hat{\beta}_j$ is 1.18, which suggests that further examination of these values would not be worthwhile.

Because the first term in the singular value decomposition stands out (the non-zero values of \hat{d}_k are 4.8, 3.7, 3.1, 2.5, 2.1, 2.1, and 1.9), the $\sum_j \hat{y}_j \hat{z}_i$ term is potentially interesting. Table 4 shows that this term is largely a contrast between, on one hand, the systems from CMU and IBM and on the other, the rest of the systems. The speakers primarily responsible for this term, Senator Hollings, Senator Kennedy, Leon Harris, Senator Dole, President Clinton, and Jim Moret, are listed in Table 5. They are listed in order of decreasing magnitude of \hat{y}_j .

Table 5. Speakers Associated With the 1997 F0 Contrast

Speaker	Number of Words	Loading \hat{y}_j
Hollings	107	7.07
Kennedy	169	3.59
Harris	896	0.92
Dole	326	1.27
Clinton	282	-1.34
Moret	47	-3.22

Compared to other speakers, the systems from CMU and IBM did poorly with the first four of these speakers and did well with the last two. We note that as in 1996, the CMU system had trouble with Senator Dole in 1997.

The 1997 system results for F1 are given in Table 6.

Table 6. 1997 Results for F1.

System	Word Error Rate	Centered WER \hat{x}_i	Contrast \hat{z}_i
bbn	19.1	-2.5	-2.4
cmu	24.2	2.6	-1.3
cu-con	27.5	5.9	-1.5
cu-htk	16.5	-5.0	1.2
dragon	24.6	3.1	5.9
ibm	19.3	-2.2	-0.1
limsi	17.4	-4.1	0.8
philips	23.0	1.4	-0.8
sri	22.3	0.8	-1.8

We see that the system word error rates are larger than those for F0 shown in Table 4 but that the system-to-system differences are not much different. There are 12 speakers. The F ratio for the $\hat{\beta}_j$ is 1.22, which suggests that further examination of these values would not be worthwhile.

The $\sum_j \hat{y}_j \hat{z}_i$ term is more interesting. The values of \hat{d}_k , which are 5.0, 2.4, 2.1, 2.0, 1.5, 1.2, and 0.8, show that the first term of the singular value decomposition stands out. We see that this term is largely a contrast between the DRAGON system and the others. The speaker that this system had trouble with is Bob Edwards who spoke 196 words, largely isolated questions. The value of \hat{y}_j for

this speaker is 3.96. Careful thought about why this is true might produce important insights.

4. SYSTEM BY FOCUS CONDITION

In this section, we consider the test data grouped by focus condition and include all seven focus conditions. The corresponding two-way table of word error rates is familiar from last year's presentation. Of interest is the fact that beyond the overall system performance, performance for focus conditions F2 (speech over telephone channels) and FX (all other speech) differentiate some systems from others. This suggests that these systems may have some advantages even though their overall performance is not the best.

The system results for 1997 and all focus conditions are given in Table 7.

Table 7. 1997 System Results for F0-FX.

System	Word error rate	Centered WER $\hat{\alpha}_i$	Contrast \hat{z}_i
bbn	19.9	-0.5	2.2
cmu	22.7	2.4	-0.6
cu-con	25.1	4.7	0.1
cu-htk	15.8	-4.6	-0.7
dragon	22.3	2.0	0.0
ibm	17.4	-3.0	0.9
limsi	17.8	-2.5	-0.6
philips	22.5	2.1	-0.3
sri	19.8	-0.6	-1.1

The corresponding focus condition results are given in Table 8.

Table 8. 1997 Focus Conditions Results for F0-FX.

Focus condition	Difficulty $\hat{\alpha}_j$	Regression $\hat{\beta}_j$	Loading $\hat{\gamma}_j$
F0	12.6	-0.36	-0.42
F1	21.5	0.16	-0.48
F2	27.1	0.63	1.95
F3	28.9	-0.04	-1.17
F4	24.0	0.21	-0.50
F5	25.5	-0.51	-1.17
FX	37.7	0.49	1.80

The values of $\hat{\alpha}_i$ and $\hat{\alpha}_j$ in Tables 7 and 8 show that CU-HTK, IBM, and LIMSI systems achieved the best performance and that focus conditions F0 and F1 are easiest and FX hardest. From the values of $\hat{\beta}_j$, we see that speech under focus conditions F2 and FX contribute most to the overall system performance.

What one cannot see from simple inspection of the table of segment-focus condition word error rates is the $\sum_j \gamma_j z_i$ term in the analysis, which characterizes performance differences more subtle than the overall system performance. The non-zero values of $\hat{\alpha}_k$ are 4.9, 3.7, 2.2, 1.5, 1.1, and 0.5. Although the first of these values does not stand out strongly, we consider the $\sum_j \gamma_j z_i$ term anyway. From the values of $\hat{\gamma}_j$ in Table 8, we see that this term differentiates the focus conditions F2 (speech from telephone channels) and FX (all other speech) from the others. From the values of \hat{z}_i in Table 7, we see that the BBN system does poorly when compared with the SRI system with speech in these two focus conditions. Conversely, we see that the BBN system does well when compared the SRI system with speech in the other focus conditions since overall, the BBN and SRI systems perform about the same. This suggests that by comparing the designs of the BBN and SRI systems, one might find the design differences responsible for this performance difference and thereby obtain ideas on system improvement.

Interestingly, the 1996 data show similar results. The system results for 1996 and all focus conditions are given in Table 9.

Table 9. 1996 System Results for F0-FX Combined.

System	Word error rate	Centered WER $\hat{\alpha}_i$	Contrast \hat{z}_i
bbn	30.4	-1.3	0.0
cmu	35.2	3.6	0.6
cu-con	34.9	3.2	0.1
cu-htk	27.8	-3.9	1.1
ibm	32.3	0.7	1.9
limsi	27.4	-4.3	-1.2
sri	33.5	1.9	-2.5

The focus condition results are given in Table 10.

Table 10. 1996 Focus Conditions Results for F0-FX.

Focus condition	Difficulty $\hat{\alpha}_j$	Regression $\hat{\beta}_j$	Loading $\hat{\gamma}_j$
F0	23.0	-0.15	-0.87
F1	30.6	-0.09	-0.35
F2	34.8	0.13	1.98
F3	28.6	0.88	-0.53
F4	37.7	0.32	0.50
F5	31.9	0.69	-0.87
FX	51.2	-0.29	1.72

The most striking difference between 1997 and 1996 is the performance exhibited by all systems. The system rankings, on the other hand, changed only moderately. Instead of comparing the values of $\hat{\beta}_j$ between 1997 and 1996, we might ignore them because the F ratio for 1996 is quite small.

The values of $\hat{\alpha}_j$ and $\hat{\gamma}_j$ tell a somewhat similar story for the 1996 data as they did for the 1997 data. We see that the contrast differentiates the focus conditions F2 and FX (as in 1997) and to a lesser extent F4 (speech under degraded acoustic conditions) from the others. We see that the SRI system (along with the LIMSI system) does better under these conditions.

5. POPULATION CONSIDERATIONS

As an alternative to viewing the broadcast news benchmark tests as a source of ideas for system improvement, one might wonder what the results imply about the superiority of one system over another. By superiority, one would usually mean that one system would perform better than another if applied to a large body of speech that one would be willing to call a population of news broadcasts. A population of news broadcasts might be all the network news shows broadcast during the last 20 years.

The big problem with inferring population performance from the 1997 benchmark test is representativeness. The ten hours of speech from which the 1997 material was selected was itself arbitrarily selected. As shown by the foregoing analysis, there are a variety of factors that affect comparative system performance. These might be present in different proportions in the ten hours than in the population of interest. Determining the effect of the lack-of-representativeness on the 1997 results on comparative performance seems difficult.

If we were to regard the 1997 test data as a random sample from some population, then we could perform a hypothesis test to see whether, in terms of this population, two systems are significantly different in their performance. The question is whether we should regard the test data as a random sample of segments, or a random sample of speakers. A random

sample of speakers is a better choice for the following reason. If we were to select a random sample of segments from a population, then we would obtain many more speakers than in the 1997 test data and thus a greater variety of speakers. Since speaker is an important determinant of performance, the segments in the 1997 test data exhibit a statistical dependence that would make the usual estimate of variance invalid and thus a hypothesis test invalid.

To test the difference in word error rates between two systems, one must realize that this difference is a ratio estimate. The numerator is the difference in total errors between the two systems and the denominator is the total number of words in the test set. In the context of random sampling, the numerator and denominator are both random variables. This case is treated by Cochran [4].

Consider the differences in the word error rate obtained from Table 1. Taking the systems two at a time and using Cochran's formula, we obtain standard deviations for the differences that range between 0.4 and 0.8. Thus, the least significant difference at the 0.05 level ranges from 0.8 to 1.6. In rough terms, differences in word error rate less than 1.2 should be regarded as perhaps only due to the peculiar selection of the 1997 test data.

6. CONCLUSIONS

At this time, no conclusion can be reached on the real value of the foregoing analysis because a careful search for connections with system features has not been done. Will listening to the segments by Senator Dole, Senator Hollings, or Bob Edwards suggest anything to system developers? Will the emergence of focus condition F2 as a type of segment that distinguishes systems suggest anything to developers? It is too early to tell.

The process of system development involves many experiments. In interpreting the results of many of these experiments, researchers may look only at the overall word error rate. The method presented here provides a way to obtain more information from such experiments. This information should speed system development.

REFERENCES

1. Fisher, W. M., "Factors Affecting Recognition Error Rate," in *Proc. Speech Recognition Workshop February 18-21, 1996*, Arden Conference Center, Harriman, NY.
2. Krishnaiah, P. R. and Yochmowitz, M. G., "Inference on the Structure of Interaction in Two-Way Classification Model," in *Handbook of Statistics, Vol. 1*, Amsterdam: North-Holland Publishing Company (1980) pp. 973-994.
3. Pallett, D. S., Fiscus, J. G., and Przybocki, M. A., "1996 Preliminary Broadcast News Benchmark Tests," in *Proc. Speech Recognition Workshop February 2-5, 1997*, Westfields International Conference Center, Chantilly, VA.
4. Cochran, W. G., *Sampling Techniques*, New York: John Wiley and Sons (1977).