

1996 PRELIMINARY BROADCAST NEWS BENCHMARK TESTS

David S. Pallett, Jonathan G. Fiscus and Mark A. Przybocki

National Institute of Standards and Technology (NIST)
Information Technology Laboratory (ITL)
Room A 216 Building 225 (Technology)
Gaithersburg, MD 20899
E-mail: dpallett@nist.gov

ABSTRACT

This paper documents use of Broadcast News test materials in DARPA-sponsored Automatic Speech Recognition (ASR) Benchmark Tests conducted late in 1996. In this year's tests, the source materials were broadened to incorporate both television and radio news broadcasts. A form of "partitioned evaluation" (PE) testing was implemented for the first time. At three sites, an additional testing protocol -- similar to that used in last year's "Dry Run" tests [1] -- was used, now termed an "Unpartitioned Evaluation" (UE). Participants in these tests included nine groups at eight sites: BBN Systems and Technologies, Cambridge University (two groups), Carnegie Mellon University, IBM, LIMSI, New York University, Rutgers University, and SRI International.

Evaluation Test Set Word Error Rates are reported for the complete evaluation test set, drawn from 4 news broadcasts (2 radio and 2 TV), and for each "Focus Condition", corresponding to seven pre-defined subsets of similarly-annotated data.

For the system with the lowest measured word error rate, the word error rate for the complete test set was 27.1%, with error rates for the focus conditions ranging from 20.3% to 46.1%.

The error rates for "found speech" vary dramatically throughout the course of a broadcast news segment, and from one segment to another, so that the test set word error rates tell only a portion of the story, and each test set -- and subset -- has its own properties. These factors are discussed at some length.

1. TRAINING AND TEST MATERIALS

The data used in this research program, and the source of the test materials, were collected by the staff of the Linguistic Data Consortium (LDC). The process of recording, digitization, and transcription this corpus is described in another paper in this Proceedings [2.]

Approximately 50 hours of recorded radio and TV newscasts were made available for system training purposes. NIST distributed these data (on sets of 20 CD-ROMs) to a community of researchers expressing tentative interest in participating in these tests, after receiving permission to do so from the LDC. In addition to the eight sites that participated in the tests, four more sites received the development test materials, but declined to

participate in the 1996 Benchmark Tests.

Additional data (amounting to a total of 20 hours) were also provided by the LDC to NIST for potential use as development and evaluation test materials. NIST collaborated with the LDC and with representatives of the DoD to review and revise the annotation and transcription of these materials. NIST also selected and distributed both a development test set and an evaluation test set. These efforts are described in another paper in this Proceedings [3].

2. TEST PARADIGM AND SCORING

Nine different research groups, at eight sites, participated in these tests -- BBN Systems and Technologies, Carnegie Mellon University, England's Cambridge University Engineering Department's "Connectionist" and "HTK" groups, IBM's T.J. Watson Laboratories, France's LIMSI group, a collaborative effort involving New York University and SRI International, Rutgers University, and SRI International. Three of these sites (BBN, CMU, and IBM) had also participated in last year's Hub 4 "Dry Run" Broadcast Materials benchmark tests.

Discussion of the properties of the systems used for these tests are contained in other papers in this Proceedings.

The "Partitioned Evaluation" test paradigm meant that it was not necessary to develop and implement usage of a "segmenter" or "chopper" software module. For the "Unpartitioned Evaluation", as in last year's Hub 4 tests, such a module was required. The three sites that participated in both the 1995 and 1996 tests (BBN, CMU, and IBM) also provided UE test results, to complement and contrast with the PE system results.

Richard Stern served to chair a Working Group including representatives of potential test participants. This Working Group defined the test protocol that was implemented as described in another paper in these Proceedings [4].

The scoring procedures for this year's evaluation followed last year's procedures with a few changes. As in last year's test, each ASR system output a "begin time" and "duration" for each recognized word. The ASR system's results were aligned and scored against time-marked "partitioned segments", using NIST's SCLITE scoring package. On average, the partitioned segments used in scoring were 54 words in length.

Before scoring, both the ASR system output and reference

transcripts were pre-filtered using orthographic transformation rules. The rules fall into four classes: (1) alternate standard spellings, (2) spelling errors in the training transcripts, (3) compound words, and (4) contractions. Rules for expansion of contractions were applied only to the hypothesis transcripts. See the discussion on "Orthographic Transformations" in another paper in these Proceedings [3].

New to this evaluation were the following.

1) Regions of overlapping speech were hand marked in the reference transcripts and automatically ignored during the scoring process.

2) Contractions were scored against their correct expanded form. This necessitated hand labeling the reference contractions to denote each contraction's correct expanded form, using context to disambiguate possible expansions.

3) Spoken word fragments in the reference transcript could match either nothing, or a hypothesized word. Since the fragment notation contains only a best guess at the sequence of letters spoken, fragments were counted as correct if the fragment's text substring matched the beginning substring of the hypothesized word. For example, the reference fragment "fr-" would match "frank" but not "find".

3. TEST RESULTS

There are numerous summary tables that can be produced to document the results of these benchmark tests. Since each partitioned segment is scored as a separate entity, and the attributes of each segment is known, consistent tabulations are readily produced, and each of these may afford opportunities for diagnostic insights. In essence, all of the NIST tabulations of test results are based on measures of word error rates (expressed as a percentage of the number of words in a test (sub)set), and these data are determined for each speaker in the test material.

Table 1 presents an example of one such report (for the ibm1 system), showing word error rates for test (sub)set word error rates for each speaker, each focus condition and (sub)set summary statistics including mean word error rates, associated standard deviations, and median word error rates. In the case of the data relating to the mean and median error rates, these operations are taken over the speaker set and are perhaps more indicative of performance over the test set population than of the test set material, since the amount of material per speaker, and the domain of the discourse, vary widely.

Each "focus condition" corresponds to a pre-defined set of transcription attributes, as described in other papers in this Proceedings.

Note that whereas the overall test set word error rate is, in this case, 32.2% for the 20,202 (scorable) word tokens, the mean word error rate is slightly higher (35.6%, with an associated standard deviation of 22.3%). The median word error rate is 29.6%. Similar observations can be made for each of the focus conditions.

The number of reference word tokens per speaker, overall, varies from 20 words to 1797 words. Note also that in some of the focus conditions, there are particularly small samples (i.e., note that the number of Bob Dole's data categorized as "under degraded acoustic conditions" involves only 7 reference words). The total number of reference words in the several focus conditions ranges from a low of 299 words in the non-native speaker focus condition to a high of 6607 in the spontaneous broadcast speech focus condition.

This attribute -- nonuniform representation of the data in the various focus conditions -- is characteristic of these "found speech" data, and must be recognized when reviewing the results.

Table 2 presents a summary report for the systems participating in the Partitioned Evaluation Benchmark Tests. The numbers tabulated are those corresponding to the related test set (or subset) word error rates. (Note, for example, that the 32.2% word error rate shown for the ibm1 system in Table 1, and discussed in a previous paragraph, also appears in this table.) These are perhaps the most frequently cited "numbers" for these tests. Table 2(a) presents data for the complete test set and each of the focus conditions, and Table 2(b) presents data, in addition, for each of the test set's component broadcasts.

For the system with the lowest measured word error rate (lims1) the word error rate for the complete test set was 27.1%, with error rates for the focus conditions ranging from 20.3% to 46.1%. Note that closely comparable results are reported for the cu-htk1 system.

In preparing the test materials [3], NIST compared and "reconciled" differences for three transcribers, and then scored the individual transcribers' transcriptions against the same "reconciled" reference strings that were used to score the automatic speech recognition systems. For the complete test set, the three individual transcribers' word error rates were 4.6%, 3.2%, and 3.2% -- almost an order of magnitude less than most automatic speech recognition systems. The lowest word error rate (0.3%) was achieved for F5, and the highest word error rate (5.4%) was achieved for FX.

In Table 2(b), note that in many, but not all, cases comparable error rates are reported for any specific system over each of the four component broadcasts. For example, for the ibm1 system, the error rate for the CNN "Morning News" material is 35.5%, 32.5% for the CSPAN "Washington Journal", and 35.3% for the NPR "The World" material. However, it is 24.9% for the NPR "Marketplace" material. For most participants, the Marketplace test materials yielded somewhat lower error rates -- possibly related to the use of Marketplace materials in the 1995 "Dry Run" tests and the researchers' greater familiarity with these broadcasts.

Figure 1 shows the error rates of Table 2(a), graphically illustrating general trends. Note that in some of the focus conditions, the relative rankings of different systems change. Consider, for example the fact for F0, F4 and F5, the cu-htk1 system (denoted as "CUHT" in this figure) has the lowest error rates, while for F1, F2 and F3 the lowest error rates are found

for the lmsil (“LIMS”) system.

Note that the two Rutgers systems differ appreciably for F4, reflecting a “bug” that was fixed for the ru2 system.

Table 3 presents a matrix tabulation of the results of NIST’s implementation of several paired-comparison significance tests, as has been provided in prior years. These significance tests are all two-tailed tests with the null hypothesis being that there is no significant performance difference between the two systems under consideration. The column at the right- and left-hand sides of the table lists abbreviations for the type of significance test. Because of the use of partitioned data, the McNemar “sentence error rate” data were in this case obtained using partitioned segments as the corresponding units. (For these tests, a “correctly recognized partitioned segment is one that is recognized without any errors.)

In prior years we printed the word “same” to indicate that the word error rates (or the sentence or partitioned segment error rates in the case of the McNemar (MN) tests) were not shown to be significantly different. In this year’s implementations, we show additional information.

Each matrix element presents data for one set of comparisons involving two systems. Within each matrix element, there are three columns of data.

The first column indicates if the test finds a significant difference at the level of $p=0.05$. If no difference is found at this level, a hyphen “-” is printed instead of the word “same” for brevity’s sake. If a difference is found at this level, this column indicates the identity of the system with the higher value on the performance statistic utilized by the particular test -- what might be regarded as the better-performing system in some sense.

The second column specifies the minimum value of p for which the test finds a significant difference at the level of p , what might be called the “exact” significance level of the test.

The third column indicates if the test finds a significant difference at the level $p=0.001$ (denoted with ***), or at the level $p=0.01$, but not $p=0.001$ (denoted with **), or at the level $p=0.05$, but not $p=0.01$ (denoted with *).

To illustrate these comparisons, consider first the matrix element corresponding to comparisons involving the bbn1 and cmu1 systems in the left hand top portion of the matrix. For three of the tests, significant differences were found at the level $p=0.05$. That is indicated with “bbn1” printed for these three tests. However, for the McNemar test, no significant difference is found at this level, and the hyphen denotes that fact. For the three tests for which significant differences were found at the $p=0.05$ level, the entry “<0.001” denotes the fact that the exact significance level is in fact less than $p=0.001$. That fact is also indicated by the printed symbols “***” in the third column. For the McNemar test, in this case, the exact significance level is shown as 1.00, the maximal possible value, corresponding roughly to insignificant differences in the performance of the two systems, on this test involving the partitioned segment error rate.

Next, consider the matrix element corresponding to comparisons involving the cu-htk1 and lmsil systems. For all of the tests, significant differences were not found at the level $p=0.05$, and that is indicated with a hyphen printed for these three tests. The exact significance levels range from a minimum of 0.180 to a maximum value of 0.562, rather large values in comparison with the results of other paired-system comparisons, indicating that the differences in performance between these two systems are certainly not pronounced.

This matrix of significance test results is applicable to the results from the entire test set, and similar tests can be applied to the results from individual focus conditions. In many cases, the results of these tests are not markedly different (i.e., consideration of the data for the entire test set yields significance test results that are frequently similar to those for any one of the focus conditions).

In any case, it is wise to bear in mind the fact that any one set of test results is just that -- one set of results for a given set of training and test data and protocols -- and the degree to which these results might indicate performance on other data is, in general, unknown.

Three of the sites (BBN, CMU, and IBM) that participated in last year’s “Dry Run” Marketplace Broadcast-based Hub 4 tests also provided results for this year’s “Unpartitioned Evaluation”. Table 3 presents the results of both the PE and UE tests. For both BBN and CMU, the differences in performance for the complete evaluation test set between the PE and UE systems are not marked. However, for IBM, a substantial difference in performance (word error rates of 28.0% for the PE system vs. 56.2% for the UE system) can be noted for the F3 focus condition (speech in the presence of background music).

4. DISCUSSION

The numbers of the preceding tables do not tell a complete story about the broadcast news data, as that data affects the instantaneous error rate. One of the most striking attributes of the broadcast news data is the rapid and frequently dramatic variability in partitioned segment word error rates throughout the broadcasts.

Figure 2 provides what might be termed a “time-line” display of the partitioned-segment word error rates vs. time for the four component broadcasts included in the 1996 Evaluation Test Set: (a) CNN “Morning News”, (b) CSPAN “Washington Journal”, (c) NPR “The World”, and (d) NPR “Marketplace”. The system from which these data were obtained for illustrative purposes is the ibm1 system. Within each broadcast, unique colors have been assigned to each speaker so as to illustrate the variability in error rate for each speaker. Note that the partitioned-segment word error rates for each speaker often vary appreciably from segment to segment throughout the broadcasts. It is particularly easy to appreciate the fact that word error rates for some speakers are markedly lower than others (e.g., see the data for Marketplace’s John Dimsdale in Figure 2(d). In contrast, for some speakers, (e.g., John McEnroe in Figure 2(c)) error rates approach or exceed 100%.

Figure 3 shows the same data as for figure 2, but in this case unique colors have been assigned to each of the Focus Conditions. Note that the distribution of materials across different broadcasts varies -- CSPAN's "Washington Journal" includes a substantial amount of spontaneous speech (the F1 focus condition), as well as telephone bandwidth speech (the F2 focus condition), and the NPR "The World" broadcast includes a substantial amount of mixed-condition speech (the FX condition). It is clear from the data of Figure 3(c) that high error rates are found with the FX data.

In these figures, the occasional gaps are due to the presence of "untestable" materials such as commercials.

Figure 4 shows the error rates in the several different focus conditions for the ibm1 system in the form of a bar graph. Each bar's width is made proportional to the number of words in each test subset. Note, for example, that the largest amount of material in any one focus condition is for the F1 spontaneous speech, while the least amount is for the F5, non-native speakers focus condition. Some general trends can be readily observed from this figure: the F0 "baseline" subset has the lowest error rates (21.6%), and next harder (30.4%) is for the spontaneous speech in F1. For F2, the higher error rates (38.9%) are probably associated both with the telephone channel and with the inclusion of spontaneous speech in this category. The amount of material in the F5 (non-native speakers) focus condition is quite small -- only 299 words -- for this test set. As noted previously, the highest error rates in any one focus condition (54.2%) are found for the FX ("all other speech", or combinations of conditions category).

Figure 5 shows the distribution of word error rates across the 1996 evaluation test set for the F0 baseline focus condition for the same system. Each speaker in this subset has been assigned a unique color, and the speakers are ordered in terms of increasing word error rate. In this graph, the width of each speaker's bar is made proportional to the number of words spoken by the speaker. Note that, while there are no obvious "outliers", there is a substantial amount of material (in fact, as Table 1 indicates, 1030 words) from the speaker with the highest error rate (~30%), Byron Miranda -- a weather-forecaster-- and this individual is responsible for a large fraction of the word errors in the F0 focus condition.

Figure 6 shows similar information for the F1 spontaneous speech focus condition. In this case, however, note the narrow bar at the right hand side of the graph (corresponding to only 44 words), but with error rates in excess of 50%, for Donna Kelly. This subset of the test material appears to be dominated by the speaker named Bill Straub.

Figure 7 includes the information shown in figures 5 and 6 for focus conditions F0 and F1, but also includes information about focus conditions F2 through FX. Note that not only do the amounts of material in each focus condition vary, but also, of course, that in some cases there are apparent "outliers" with unusually high error rates.

These figures are intended to graphically underscore what should be obvious -- that in working with "found speech", the

properties of any one set or subset of data may differ dramatically from other sets. It should not be surprising, therefore, to find marked differences between any two test sets.

BBN provided NIST with data for this year's development and evaluation test sets, using the bbn1 system. The word error rates for the F0 focus condition are shown in Figure 8. Note that for the development test set, the dominant largest single block of material (1092 words) is for David Brancaccio, for whom error rates of 14.3% are found. In contrast, in the evaluation test set, as previously noted, the largest contribution (1030 words) is that due to Byron Miranda, with error rates of 32.8%. For these results, comparisons involving the test set mean or median word error rate (over the sets of speakers) may be more informative than the total test set word error rates because of the nonuniform distributions of source materials.

Figure 9 presents a comparison of Partitioned and Unpartitioned Test Word Error Rates for each focus condition for the two IBM systems: ibm1 (PE) and ibm2 (UE). Note that, in general and for these two systems, error rates are higher for the Unpartitioned Evaluation than for the Partitioned Evaluation.

In preliminary exploratory analyses at NIST [5], the results of the tests have been represented in the form of a two-way table with partitioned segment word error rates. The segments were assigned to rows, the systems to columns, and the word error rates to cells (the intersections of the rows and columns). The use of a transformation technique, averaging, and "centering" the averages suggests that the systems participating in these tests "seem to break into three groups, the best... the next best... and the others". (The Rutgers' systems' data were excluded from these studies.)

These exploratory analyses also suggest that many of the observed system differences are due to differences in dealing with long segments. Table 5 shows the results for focus conditions F0 and F1, when scoring is performed and results tabulated for three subsets of the data for each focus condition: (1) segments with fewer than 10 words, (2) segments with 10 to 49 words, and (3) segments with 50 or more words.

Note that, in many (but not all) cases, word error rates are lower for the longer segments (e.g., note that for the htk1 system, for F0, the word error rate ranges from 25.6% for the short segments to 18.9% for the long segments). For F0, exceptional cases include cu-con1 and limsi1 -- each of which have lower error rates for shorter segments than for longer segments. Although the validity of these generalizations may be limited by small-sample effects, the same general effect is noted for both F0 and F1.

For the different length segment subsets, performance differences between systems are interesting: for the short segments in F0, the cu-con1 system has the lowest error rate (18.6%), and for the corresponding F1 segments, the lowest error rate is found for the bbn1 system (36.6%). For the mid-length segments (10-49 words), markedly lower word error rates are found for the cu-htk1 system than for other systems, and for the F1 data, the cu-htk1 and limsi systems have comparable low word error rates. For the long segments, the

lowest word error rate for the F0 data (18.9%) is found for the cu-htk1 system, and for the F1 data, the lowest word error rate (24.6%) is found for the limsi1 system.

This dependence of relative system performance on segment length does not appear to be just the result of the random selection of segments. Implementations of NIST's paired-comparison statistical significance tests on these subsetted results indicate, in general, greater significance to individual sites' paired comparison tests with increasing segment length.

It seems likely that these differences in different systems' abilities to deal with long segments may be due to differences in acoustic and/or linguistic segmentation.

It also becomes evident that there are differing numbers of segments for which particular systems had the best performance, and that "there may be some types of segments" for which the grouping based on averages does not apply. There is evidence to suggest that the easiest segments and the hardest segments do little to distinguish the systems. Performance (across systems) is most variable for the moderately challenging segments. Other interesting questions include consideration of whether "the differences in system performance"... "would be observed for a much larger selection of news broadcasts". Continuations of these preliminary studies may yield additional insights, especially when other considerations of properties of the data are included.

5. ACKNOWLEDGMENTS

The community is greatly indebted to the staff of the LDC, especially Rebecca Finch for obtaining rights to use data from numerous "Broadcast News" sources. The authors acknowledge with thanks the assistance of Dave Graff at the LDC for his role in collecting and coordinating the annotation and transcription of the training and test materials at the LDC.

Shirley Ramsey, a DoD employee, participated in the test data annotation and transcription efforts on-site at NIST. It was a pleasure to have her assistance and to work with her.

AT NIST, John Garofolo and Jon Fiscus deserve special credit for their role in working with Dave Graff and George Doddington in developing the annotation and transcription convention. Greg Sanders, a new member of our group at NIST, participated in the test data annotation and transcription efforts, and his assistance and cooperation are greatly appreciated. Bill Fisher participated in selecting, and reconciliation of the annotations and transcriptions for, the evaluation test data. Alvin Martin participated in revision of the implementation of the paired-comparison significance tests together with Jon Fiscus. Jon Fiscus once again deserves special thanks for scoring all of the results expeditiously. Mark Przybocki has provided much assistance in too many ways to enumerate -- but especially in processing data for distribution in many of our benchmark tests. It is a pleasure to acknowledge helpful discussions with Walter Liggett, a member of NIST's Statistical Engineering Division,

in reviewing some properties of the test results. Bruce Lund assisted with preliminary preparations to prepare the Proceeding document. And finally, Kathy Gallo has helped in many ways, but especially in making sure that the training and test materials get sent out, on time and to the appropriate recipients.

NOTICE

The views expressed in this paper are those of the author(s). The results are for local, system-developer implemented tests. NIST's role was one that involved working with the LDC in processing LDC-provided training and potential test materials, selecting and defining reference annotation and transcription files for the tests, developing and implementing scoring software, and uniformly scoring and tabulating results. The views of the author(s), and these results, are not to be construed or represented as endorsements of any systems, or as official findings on the part of NIST, DARPA, or the U.S. Government.

REFERENCES

1. Pallett, D.S. et al., "1995 Hub-4 "Dry Run" Broadcast Materials Benchmark Tests", in *Proc. Speech Recognition Workshop February 18-21, 1996*, Arden Conference Center, Harriman, NY.
2. Graff, D., et al., "The 1996 Broadcast News Speech and Language-Model Corpus", in this Proceedings.
3. Garofolo, J.S., Fiscus, J.G., and Fisher, W.M., "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora", in this Proceedings.
4. Stern, R.M., "Specification of the 1996 Broadcast News Evaluation", in this Proceedings.
5. Liggett, W., "Exploratory Analysis of 1996 Broadcast News Benchmark Tests", private communication to D.S. Pallett, 27 January, 1997.

System: ibm1

Overall -> All Speech from Focus Conditions F0-F5 and FX.
 Baseline Broadcast Speech -> F0: Speech that is directed to the general broadcast audience, and that is recorded in a quiet studio environment.
 Spontaneous Broadcast Speech -> F1: Speech that is directed to one or more human conversational partners, recorded in a quiet studio environment.
 Speech Over Telephone Channels -> F2: Speech that is collected over reduced-bandwidth conditions, such as local or long distance telephony.
 Speech in the Presence of Background Music -> F3: Speech that satisfies the attributes of F0 or F1, except that it is broadcast with additive background music.
 Speech Under Degraded Acoustic Conditions -> F4: Speech that satisfies the attributes of F0 or F1, except that it is broadcast with additive background noise.
 Speech from Non-Native Speakers -> F5: Speech that satisfies the attributes of F0, except that it is spoken by non-native speakers of American English.
 All other speech -> FX: Speech which satisfies none of the F0-F5 Focus conditions.

SPKR	1996 Hub4 Focus Conditions															
	Overall		Baseline Broadcast Speech		Spontaneous Broadcast Speech		Speech Over Telephone Channels		Speech in the Presence of Background Music		Speech Under Degraded Acoustic Conditions		Speech from Non-Native Speakers		All other speech	
	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE
leon_harris	[1176]	38.9	[408]	29.9	[447]	48.5			[65]	30.8	[256]	38.3				
steve_hurst	[608]	26.6									[608]	26.6				
donna_kelly	[630]	23.2	[299]	25.4	[44]	59.1			[24]	8.3	[246]	15.0			[17]	29.4
byron_miranda	[1197]	35.2	[1030]	33.2	[42]	35.7			[125]	51.2						
kay_bailey_hutchison	[427]	16.6			[423]	16.1					[4]	75.0				
bill_richardson	[348]	20.7			[348]	20.7										
maureena_colby	[363]	98.9									[363]	98.9				
susan_swain	[821]	22.2			[821]	22.2										
file2_johndoe002	[269]	47.2					[269]	47.2								
bill_straub	[1797]	33.4			[1797]	33.4										
steven_thomma	[953]	36.0			[944]	35.9					[9]	44.4				
file2_johndoe003	[418]	34.4					[418]	34.4								
file2_janedoe001	[276]	26.4					[276]	26.4								
file2_johndoe004	[172]	45.3					[172]	45.3								
file2_johndoe005	[200]	39.5					[200]	39.5								
file2_johndoe006	[79]	39.2					[79]	39.2								
file2_johndoe007	[188]	51.1					[188]	51.1								
file2_janedoe002	[66]	7.6									[66]	7.6				
bill_clinton	[301]	22.9	[261]	19.9							[40]	42.5				
bob_dole	[288]	24.3	[281]	23.1							[7]	71.4				
mary_ambrose	[878]	23.7	[551]	19.8	[183]	34.4			[144]	25.0						
lisa_mullins	[980]	20.5	[422]	19.2	[344]	22.1			[199]	21.1	[15]	13.3				
tariq_abdul_nagib	[523]	44.2													[523]	44.2
file3_johndoe001	[146]	31.5	[20]	0.0					[126]	36.5						
karin_henrikson	[285]	31.6													[285]	31.6
ignacio_besaudi	[481]	40.1													[481]	40.1
kimberly_dozier	[378]	10.6	[378]	10.6												
zafira_bas	[72]	95.8													[72]	95.8
renaht_ahkturin	[20]	20.0													[20]	20.0
charles_scanlon	[75]	29.3														
slave_pashovski	[249]	96.8														
boris_maximov	[331]	64.0														
elena_ppd	[157]	61.1														
john_mcenroe	[25]	104.0														
bud_collins	[342]	19.3			[342]	19.3										

Table 1 Example tabulation of word error rates for test (sub)set for each speaker, each focus condition and (sub)set summary statistics including mean word error rates, associated standard deviations, and median word error rates. System: ibm1.

SPKR	Overall		Baseline Broadcast Speech		Spontaneous Broadcast Speech		Speech Over Telephone Channels		Speech in the Presence of Background Music		Speech Under Degraded Acoustic Conditions		Speech from Non-Native Speakers		All other speech	
	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE	#Wrd	%WE
file4_janedoe001	[110]	15.5														
david_branaccio	[1315]	18.6	[634]	11.0	[160]	28.7			[110]	15.5					[14]	28.6
will_durst	[496]	29.2	[452]	28.1					[500]	23.4	[7]	100.0				
john_dimsdale	[302]	7.9	[302]	7.9					[44]	40.9						
philip_boroff	[161]	13.0	[161]	13.0												
barbara_boxer	[41]	31.7					[41]	31.7								
joanne_miles	[107]	37.4					[107]	37.4								
david_johnson	[471]	34.2			[457]	33.9			[14]	42.9						
george_lewinski	[132]	16.7	[104]	9.6					[28]	42.9						
paul_hawkins	[248]	23.8	[229]	19.7					[14]	50.0					[5]	140.0
file4_johndoe001	[59]	23.7									[59]	23.7				
wolfgang_odnall	[46]	26.1													[46]	26.1
odmir_moslow	[52]	76.9													[52]	76.9
john_parker	[248]	35.1											[224]	31.3	[24]	70.8
fritz_ferber	[581]	26.5	[463]	24.4					[24]	41.7	[94]	33.0				
raphaela_pope	[168]	41.7			[160]	38.7					[8]	100.0				
claudia_sloan	[51]	43.1									[51]	43.1				
sam_louis	[38]	23.7			[38]	23.7										
lee_zasloff	[30]	20.0			[30]	20.0										
christina_zelaya	[27]	29.6			[27]	29.6										
Set Sum/Avg	[20202]	32.2	[5995]	21.6	[6607]	30.4	[1750]	38.9	[1417]	28.0	[1833]	42.2	[299]	30.8	[2301]	54.2
Mean	[367]	35.6	[374]	18.4	[388]	30.7	[194]	39.2	[109]	33.1	[122]	48.9	[149]	30.3	[153]	62.0
StdDev	[377]	22.3	[237]	9.1	[451]	11.3	[115]	7.8	[131]	13.5	[174]	32.3	[105]	1.4	[177]	35.3
Median	[269]	29.6	[340]	19.7	[342]	29.6	[188]	39.2	[65]	36.5	[51]	42.5	[149]	30.3	[52]	61.1

Table 1(Continued) Example tabulation of word error rates for test (sub)set for each speaker, each focus condition and (sub)set summary statistics including mean word error rates, associated standard deviations, and median word error rates. System: ibm1.

DARPA CSR 1996 Broadcast News Hub-4 Benchmark Test								
Word Error Rate Summary for the Complete Test Set and Focus Condition								
System	Complete Test	F0	F1	F2	F3	F4	F5	FX
bbn1	30.2	21.6	29.5	32.7	23.3	38.4	31.8	49.9
cmu1	34.9	25.8	32.1	38.6	36.6	43.7	36.5	55.8
cu-con1	34.7	25.8	33.5	40.4	33.4	39.3	40.5	53.1
cu-htk1	27.5	18.7	26.5	33.1	23.6	29.1	21.7	51.0
ibm1	32.2	21.6	30.4	38.9	28.0	42.2	30.8	54.2
limsil	27.1	20.8	26.0	27.1	20.3	33.3	27.8	46.1
nyu1	33.0	26.0	32.5	32.6	34.2	38.4	31.1	48.1
ru1	56.1	43.0	51.7	74.6	50.0	81.6	54.8	72.1
ru2	53.8	42.7	51.9	72.9	50.0	59.2	54.8	71.9
sril	33.3	26.4	33.0	31.7	34.7	38.5	34.4	48.3

F0 -> Baseline Broadcast Speech
 F1 -> Spontaneous Broadcast Speech
 F2 -> Speech Over Telephone Channels
 F3 -> Speech in the Presence of Background Music
 F4 -> Speech Under Degraded Acoustic Conditions
 F5 -> Speech from Non-Native Speakers
 FX -> All other speech

Table 2(a) DARPA CSR 1996 Partitioned Evaluation Broadcast News Hub-4 Benchmark Test: Word Error Rate Summary for the Complete Test Set and Focus Conditions

DARPA CSR 1996 Broadcast News Hub-4 Benchmark Test					
Word Error Rate Summary for the Complete Test Set and by Broadcast					
System	Complete Test	CNN Morning News	CSP Wash. Journal	NPR The World	NPR Marketplace
bbn1	30.2	32.8	29.8	33.1	24.8
cmu1	34.9	37.1	35.7	36.9	29.8
cu-con1	34.7	35.0	36.4	36.2	30.5
cu-htk1	27.5	28.4	27.7	32.0	21.5
ibm1	32.2	35.5	32.5	35.3	24.9
limsil	27.1	29.7	25.6	30.5	23.0
nyu1	33.0	34.2	32.2	35.6	29.9
ru1	56.1	60.6	60.3	52.9	49.6
ru2	53.8	53.5	60.5	52.2	47.7
sril	33.3	35.0	32.0	35.9	30.6

Table 2(b) DARPA CSR 1996 Partitioned Evaluation Broadcast News Hub-4 Benchmark Test: Word Error Rate Summary for the Complete Test Set and by Broadcast

Composite Report of All Significance Tests
For the DARPA CSR 1996 Broadcast News Hub-4 Partitioned Evaluation Benchmark Test

		Test Name										Abbrev.	
		Matched Pair Sentence Segment (Word Error)										MP	
		Signed Paired Comparison (Speaker Word Error Rate (%))										SI	
		Wilcoxon Signed Rank (Speaker Word Error Rate (%))										WI	
		McNemar (Partition Segment Error)										MN	
Test Abbrev.		hbhl	cmul	cu-conl	cu-htkl	ibhl	lmsil	nyul	rul	ru2	sril	Test Abbrev.	
MP	hbhl	hbhl <0.001 ***	hbhl <0.001 ***	cu-htkl <0.001 ***	hbhl <0.001 ***	lmsil <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	MP	
SI	hbhl	hbhl <0.001 ***	hbhl <0.001 ***	cu-htkl 0.001 **	hbhl 0.032 *	lmsil <0.001 ***	hbhl 0.032 *	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	SI	
WI	hbhl	hbhl <0.001 ***	hbhl <0.001 ***	cu-htkl 0.001 **	hbhl 0.020 *	lmsil <0.001 ***	hbhl 0.016 *	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	hbhl 0.022 *	WI	
MN		- 1.000	- 0.667	- 0.112	- 0.589	- 0.395	- 0.267	hbhl <0.001 ***	hbhl <0.001 ***	hbhl <0.001 ***	- 0.187	MN	
MP	cmul		- 0.589	cu-htkl <0.001 ***	ibhl <0.001 ***	lmsil <0.001 ***	nyul 0.002 **	cmul <0.001 ***	cmul <0.001 ***	sril 0.014 *		MP	
SI			- 1.000	cu-htkl <0.001 ***	ibhl 0.043 *	lmsil <0.001 ***	nyul 0.032 *	cmul <0.001 ***	cmul <0.001 ***	sril 0.032 *		SI	
WI			- 0.904	cu-htkl <0.001 ***	ibhl 0.043 *	lmsil <0.001 ***	nyul 0.024 *	cmul <0.001 ***	cmul <0.001 ***	sril 0.041 *		WI	
MN			- 0.857	- 0.055	- 0.459	- 0.250	- 0.327	cmul <0.001 ***	cmul <0.001 ***	- 0.230		MN	
MP	cu-conl			cu-htkl <0.001 ***	ibhl <0.001 ***	lmsil <0.001 ***	nyul <0.001 ***	cu-conl <0.001 ***	cu-conl <0.001 ***	sril 0.007 **		MP	
SI				cu-htkl <0.001 ***	ibhl 0.016 *	lmsil <0.001 ***	- 0.180	cu-conl <0.001 ***	cu-conl <0.001 ***	- 0.107		SI	
WI				cu-htkl <0.001 ***	ibhl 0.029 *	lmsil <0.001 ***	- 0.095	cu-conl <0.001 ***	cu-conl <0.001 ***	- 0.105		WI	
MN				cu-htkl 0.026 **	- 0.110	- 0.110	- 0.575	cu-conl <0.001 ***	cu-conl <0.001 ***	- 0.447		MN	
MP	cu-htkl			cu-htkl <0.001 ***	ibhl <0.001 ***	lmsil <0.001 ***	nyul <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	MP	
SI				cu-htkl <0.001 ***	ibhl 0.016 *	lmsil <0.001 ***	- 0.180	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	SI	
WI				cu-htkl <0.001 ***	ibhl 0.029 *	lmsil <0.001 ***	- 0.095	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	WI	
MN				cu-htkl 0.026 **	- 0.110	- 0.110	- 0.575	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	cu-htkl <0.001 ***	MN	
MP	ibhl				cu-htkl <0.001 ***	lmsil <0.001 ***	- 0.197	ibhl <0.001 ***	ibhl <0.001 ***	- 0.056		MP	
SI					cu-htkl <0.001 ***	lmsil <0.001 ***	- 0.285	ibhl <0.001 ***	ibhl <0.001 ***	- 0.180		SI	
WI					cu-htkl <0.001 ***	lmsil <0.001 ***	- 0.795	ibhl <0.001 ***	ibhl <0.001 ***	- 0.582		WI	
MN					cu-htkl <0.001 ***	lmsil <0.001 ***	- 0.865	ibhl <0.001 ***	ibhl <0.001 ***	ibhl 0.032 *		MN	
MP	lmsil					lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	MP	
SI						lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	SI	
WI						lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	WI	
MN						lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil <0.001 ***	lmsil 0.011 *	MN	
MP	nyul						nyul <0.001 ***	nyul <0.001 ***	nyul <0.001 ***	nyul 0.019 *		MP	
SI							nyul <0.001 ***	nyul <0.001 ***	nyul <0.001 ***	- 0.596		SI	
WI							nyul <0.001 ***	nyul <0.001 ***	nyul <0.001 ***	- 0.697		WI	
MN							nyul 0.007 **	nyul 0.007 **	nyul 0.007 **	- 1.000		MN	
MP	ru1								ru2 0.015 *	sril <0.001 ***		MP	
SI									- 0.285	sril <0.001 ***		SI	
WI									ru2 0.039 *	sril <0.001 ***		WI	
MN									- 1.000	sril 0.011 *		MN	
MP	ru2									sril <0.001 ***		MP	
SI										sril <0.001 ***		SI	
WI										sril <0.001 ***		WI	
MN										sril 0.011 *		MN	
MP	sril									sril <0.001 ***		MP	
SI										sril <0.001 ***		SI	
WI										sril <0.001 ***		WI	
MN										sril 0.011 *		MN	

These significance tests are all two-tailed tests with null the hypothesis that there is no performance difference between the two systems.

The first column indicates if the test finds a significant difference at the level of p=0.05. It consists of '-' if no difference is found at this significance level. If a difference at this level is found, this column indicates the system with the higher value on the performance statistic utilized by the particular test.

The second column specifies the minimum value of p for which the test finds a significant difference at the level of p.

The third column indicates if the test finds a significant difference at the level of p=0.001 (****), at the level of p=0.01, but not p=0.001 (***), or at the level of p=0.05, but not p=0.01 (**).

A test finds significance at level p if, assuming the null hypothesis, the probability of the test statistic having a value at least as extreme as that actually found, is no more than p.

Table 3 Complete significance test summary matrix for Partitioned Evaluation.

DARPA CSR 1996 Broadcast News Hub-4 Benchmark Test								
Word Error Rate Summary for the Complete Test Set and Focus Condition								
System	Complete Test	F0	F1	F2	F3	F4	F5	FX
bbn1 PE	30.2	21.6	29.5	32.7	23.3	38.4	31.8	49.9
bbn2 UE	31.8	22.8	31.6	34.3	27.1	38.8	38.1	50.8
cmu1 PE	34.9	25.8	32.1	38.6	36.6	43.7	36.5	55.8
cmu2 UE	35.9	24.7	33.1	39.1	48.4	42.1	35.5	58.3
ibm1 PE	32.2	21.6	30.4	38.9	28.0	42.2	30.8	54.2
ibm2 UE	38.9	26.8	36.8	42.4	56.2	43.0	34.1	60.7

F0 -> Baseline Broadcast Speech
 F1 -> Spontaneous Broadcast Speech
 F2 -> Speech Over Telephone Channels
 F3 -> Speech in the Presence of Background Music
 F4 -> Speech Under Degraded Acoustic Conditions
 F5 -> Speech from Non-Native Speakers
 FX -> All other speech

Table 4(a) Comparison of Partitioned Evaluation and Unpartitioned Evaluation systems tests. Word Error Rate Summary for the Complete Test Set and Focus Conditions

DARPA CSR 1996 Broadcast News Hub-4 Benchmark Test					
Word Error Rate Summary for the Complete Test Set and by Broadcast					
System	Complete Test	CNN Morning News	CSP Wash. Journal	NPR The World	NPR Marketplace
bbn1 PE	30.2	32.8	29.8	33.1	24.8
bbn2 UE	31.8	32.7	31.3	34.5	28.8
cmu1 PE	34.9	37.1	35.7	36.9	29.8
cmu2 UE	35.9	37.3	34.4	41.3	30.7
ibm1 PE	32.2	35.5	32.5	35.3	24.9
ibm2 UE	38.9	39.1	36.8	42.9	37.2

Table 4(b) Comparison of Partitioned Evaluation and Unpartitioned Evaluation systems tests. Word Error Rate Summary for the Complete Test Set and by Broadcast

Word Error Rates For Focus Conditions F0 and F1							
System	F0 and F1	F0 Focus Condition (Baseline Speech)			F1 Focus Condition (Spontaneous Speech)		
		Segment Word Lengths			Segment Word Lengths		
	ALL Seg	0-9	10-49	50or>	0-9	10-49	50or>
bbn1	25.8	23.3	22.1	21.5	36.6	33.2	28.2
cmul	29.1	34.9	22.5	26.6	39.6	32.9	31.6
cu-con1	29.8	18.6	24.9	26.2	42.7	34.1	32.9
cu-htk1	22.8	25.6	17.4	18.9	37.9	28.0	25.6
ibm1	26.2	24.4	24.1	20.9	40.5	33.1	29.3
limsil	23.5	19.8	20.2	21.0	38.8	29.4	24.6
nyul	29.4	31.4	25.0	26.1	45.4	36.9	30.7
ru1	47.6	37.2	49.3	41.4	70.5	50.8	51.1
ru2	47.5	37.2	44.4	42.3	70.5	50.2	51.5
sril	29.9	31.4	24.8	26.7	44.9	37.3	31.4

Table 5 Word Error Rates for Focus Conditions F0 and F1, partitioned into different segment length subsets.

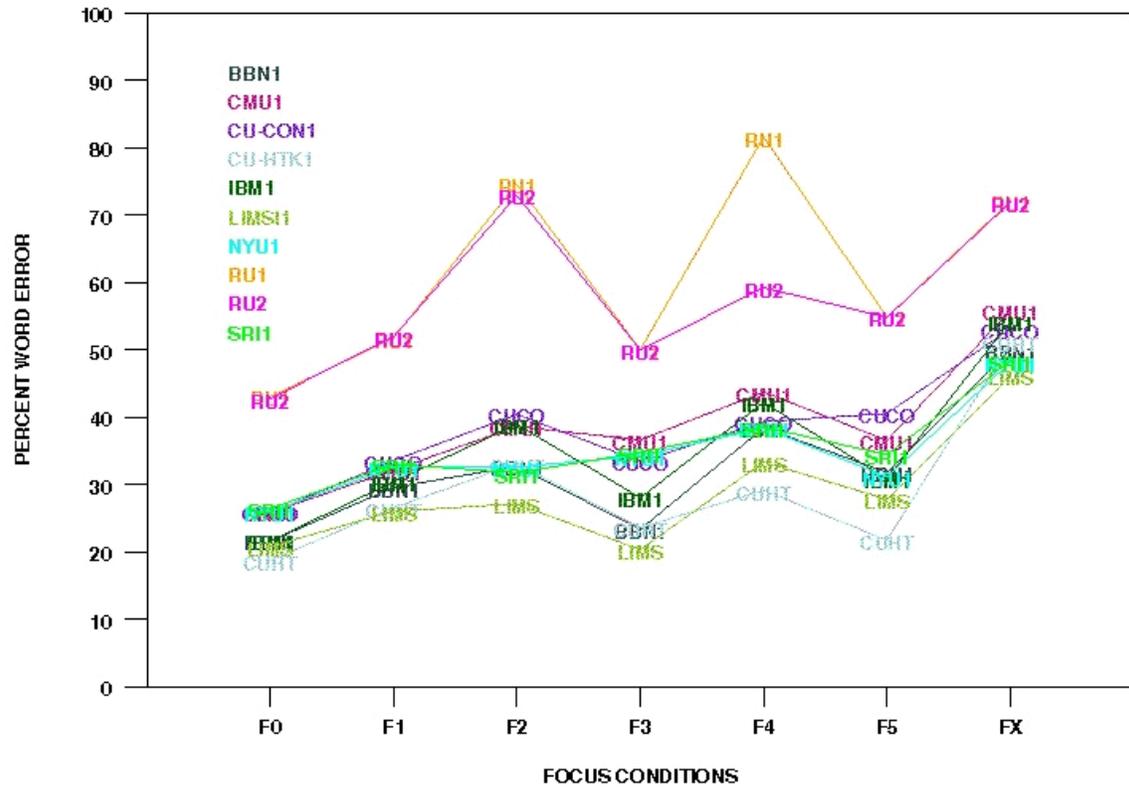


Figure 1: Site Comparison

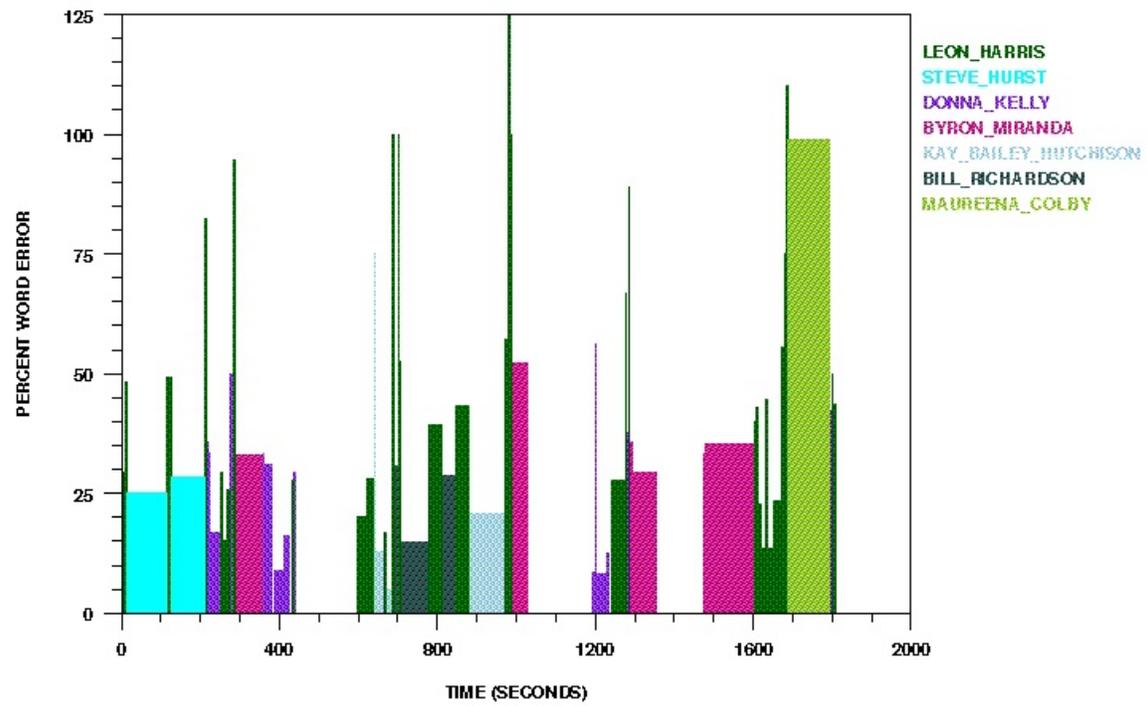


Figure 2(a): CNN Morning News

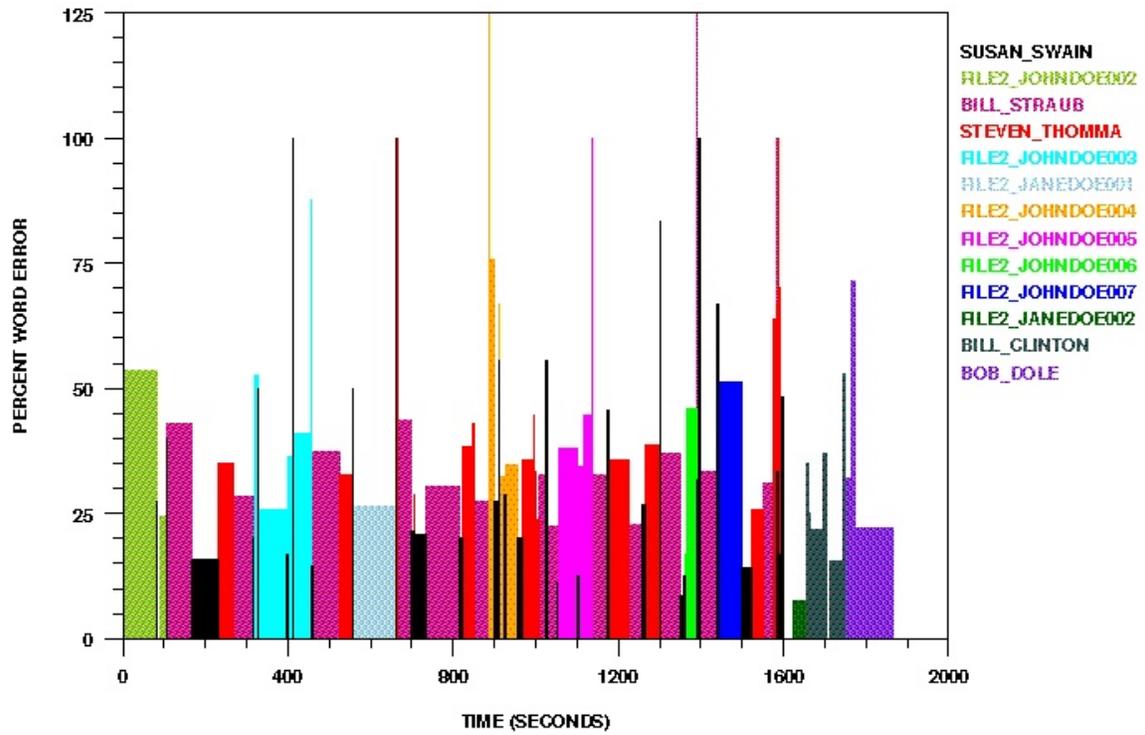


Figure 2(b): CSPAN Washington Journal

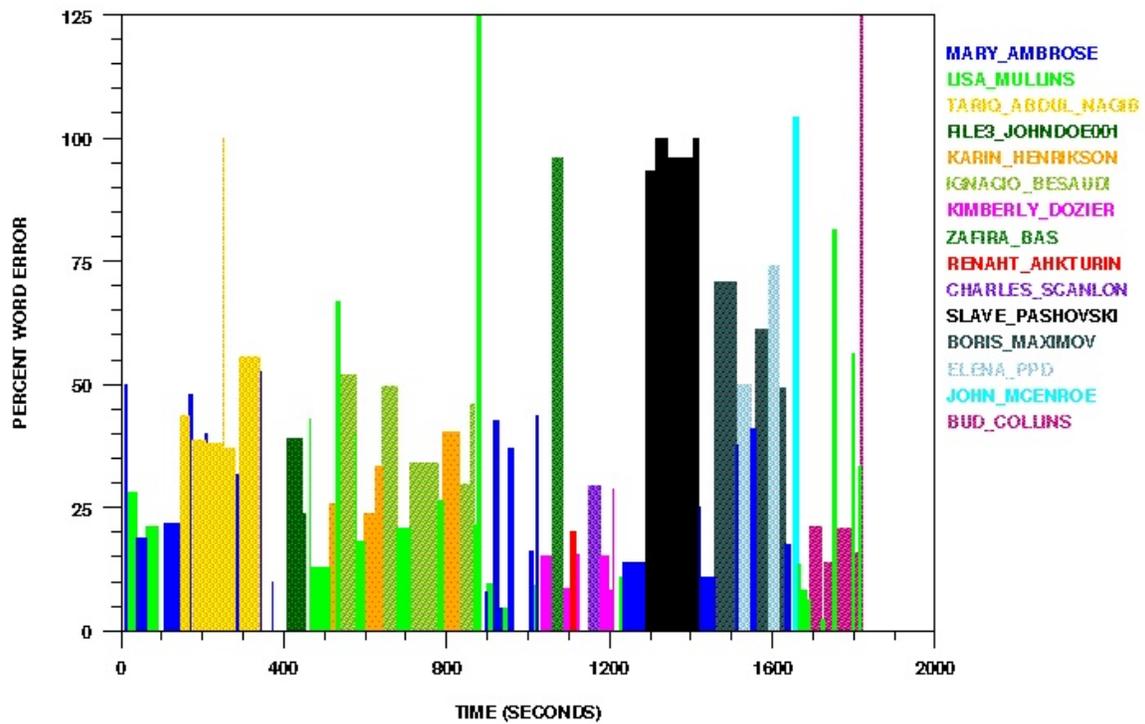


Figure 2(c): NPR The World

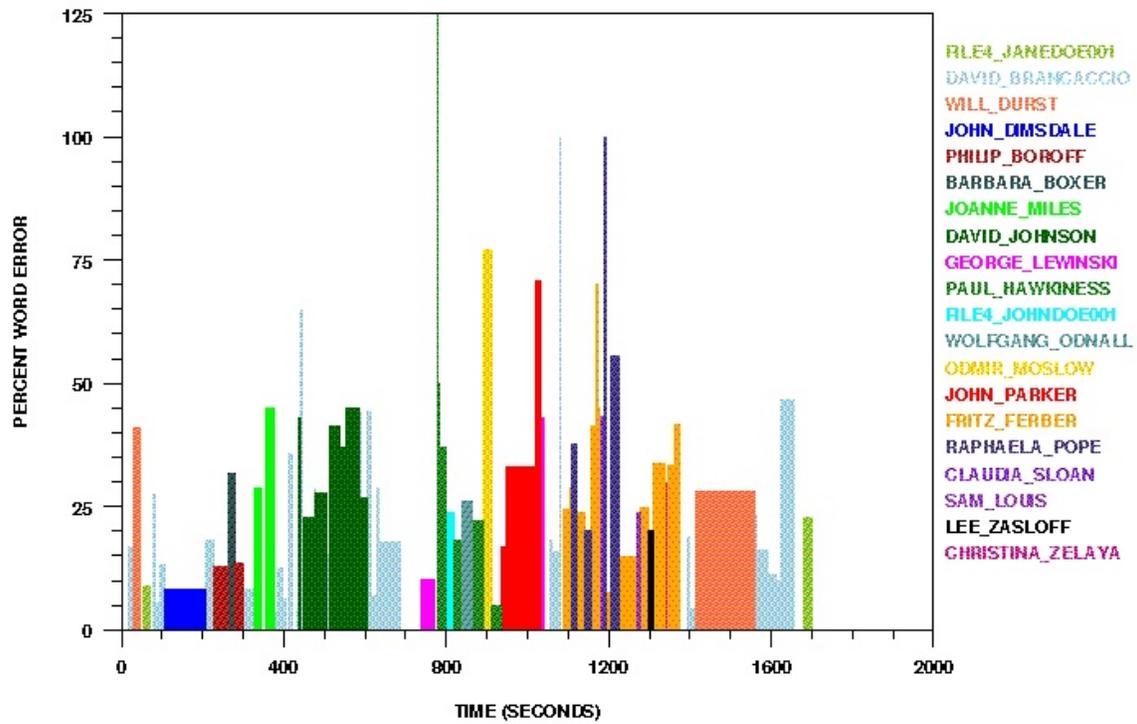


Figure 2(d): PRI Marketplace

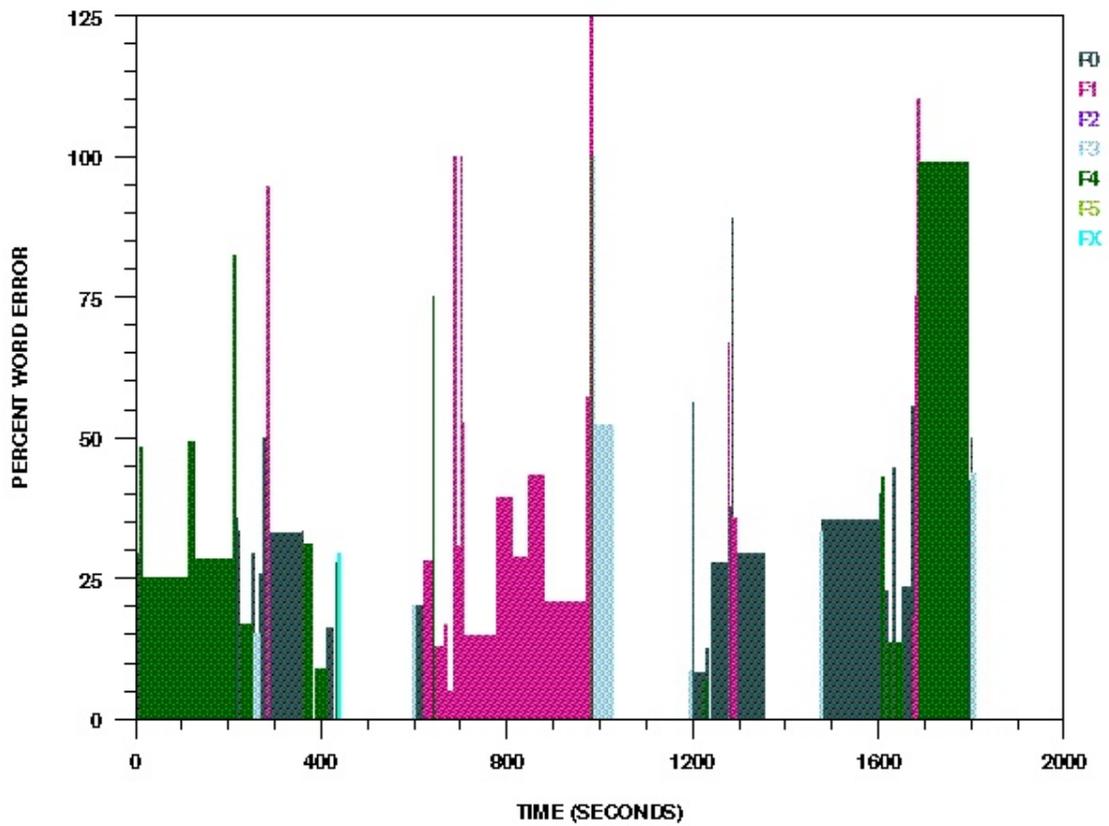


Figure 3(a): CNN Morning News

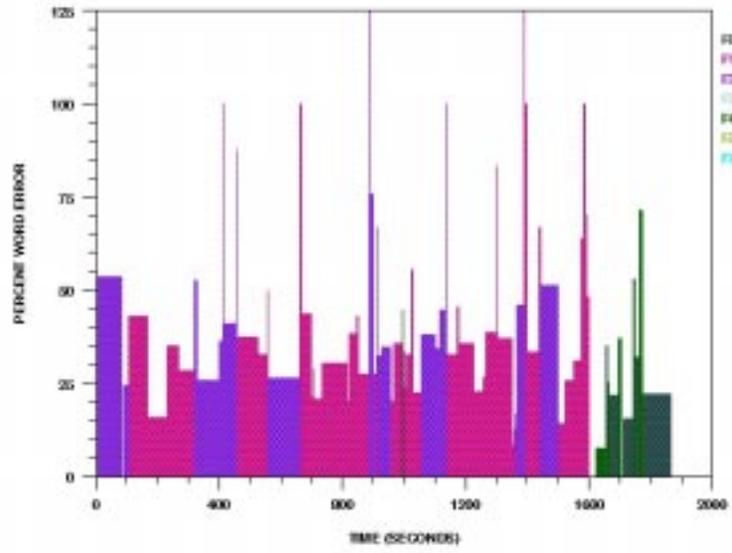


Figure 3(b): CSPAN Washington Journal

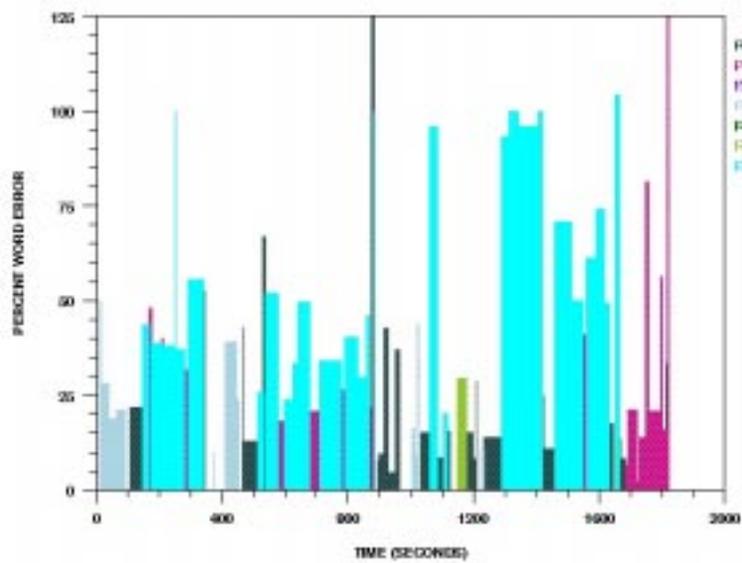


Figure 3(c): NPR The World

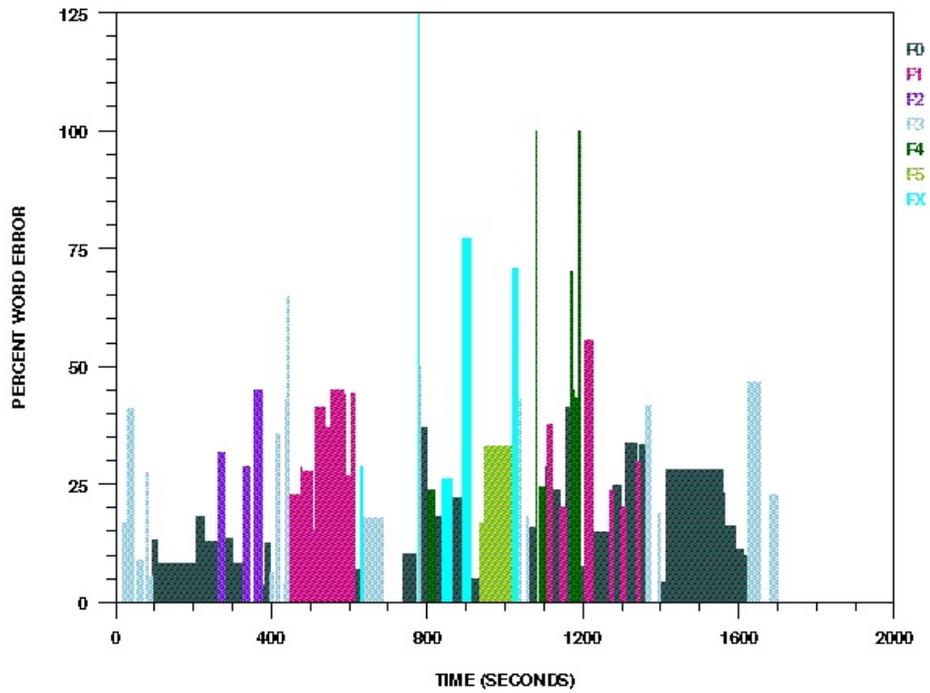


Figure 3(d): PRI Marketplace

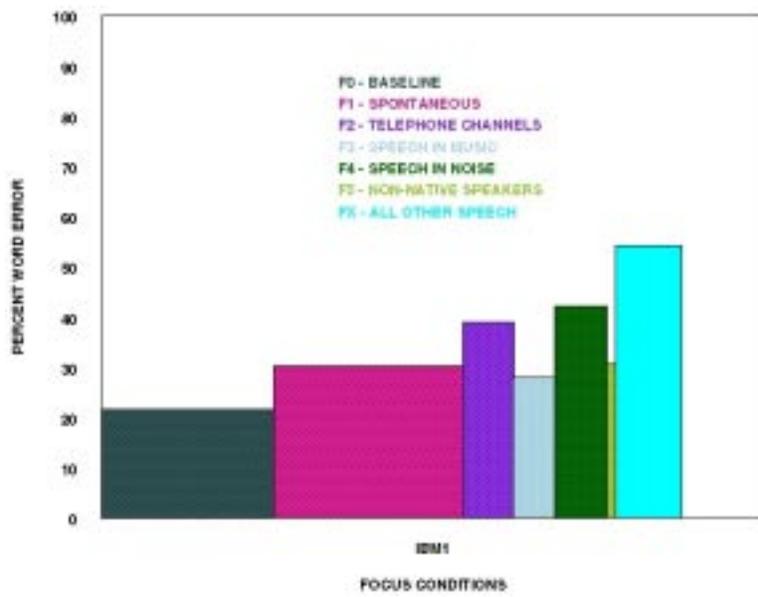


Figure 4: IBM1 - Focus Conditions

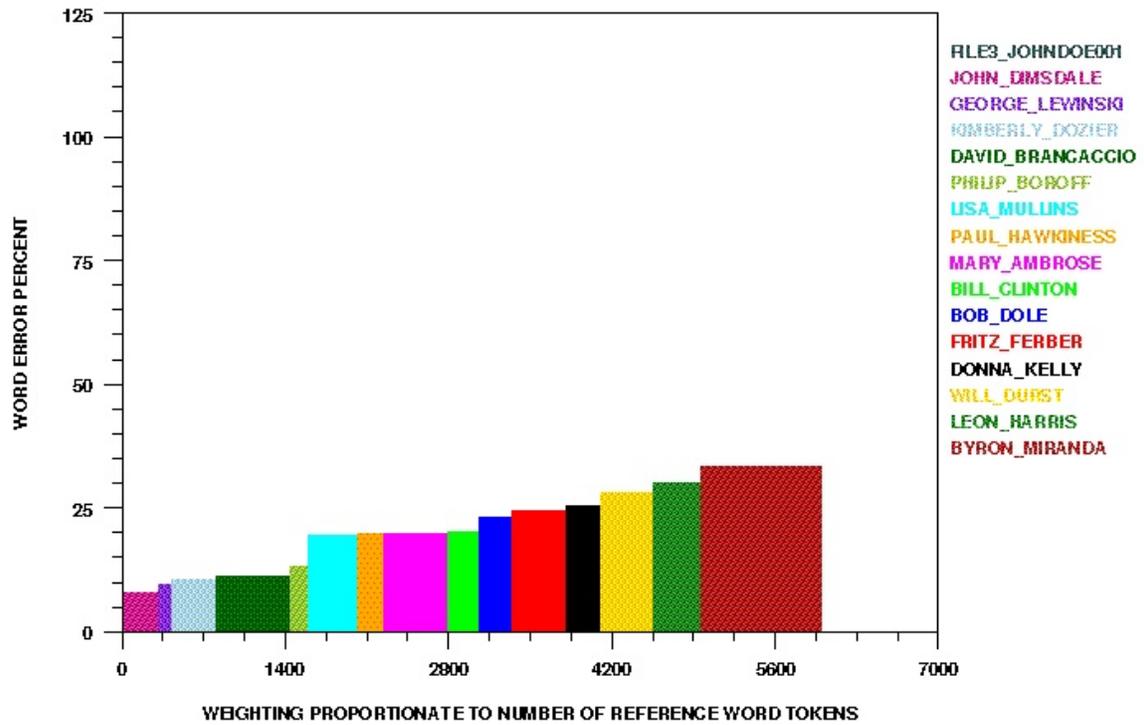


Figure 5: IBM1 - DARPA 1996 F0 (Baseline) Focus Condition

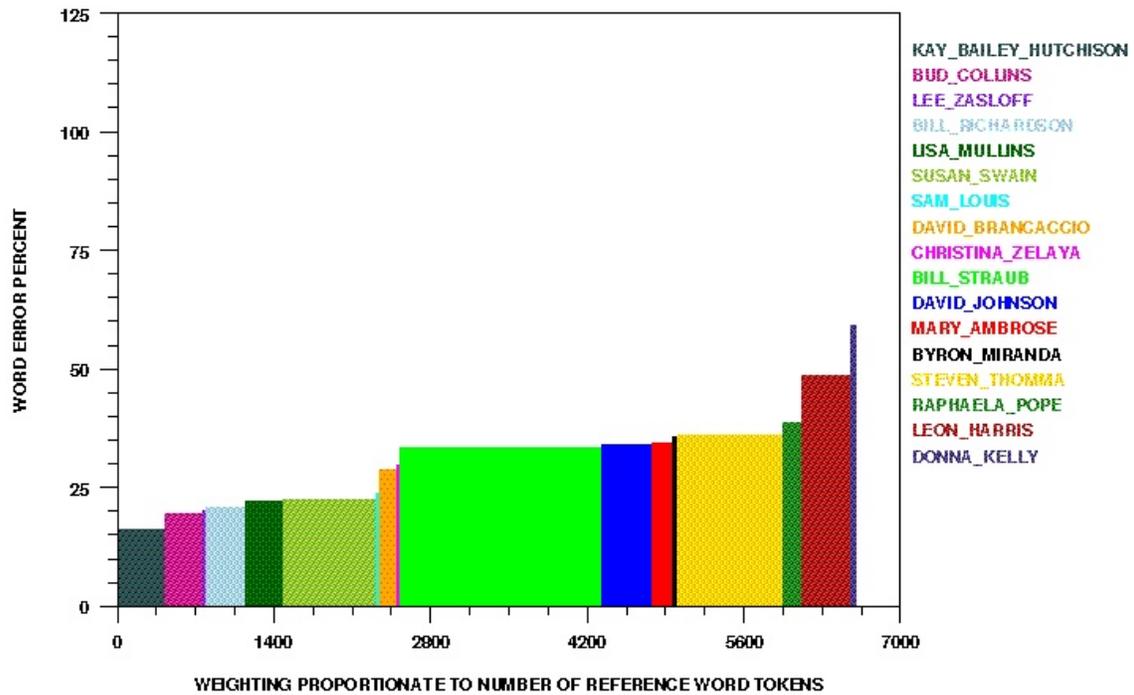


Figure 6: IBM1-DARPA 1996 F1 (Spontaneous) Focus Condition

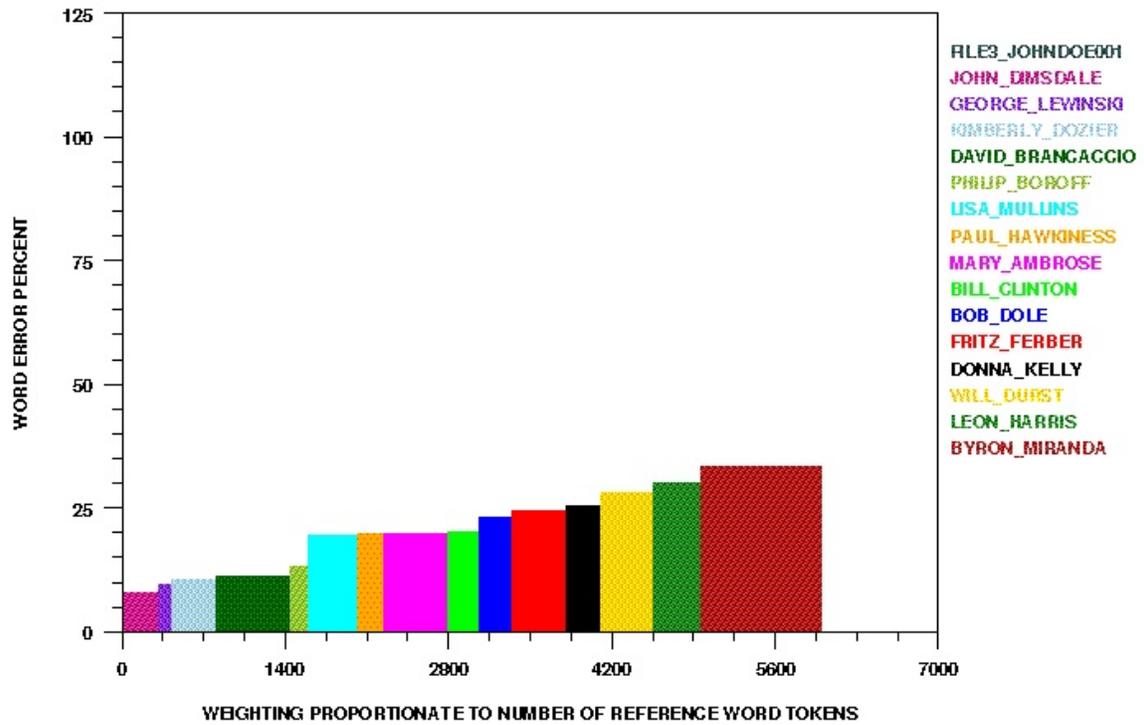


Figure 7(a): IBM1-DARPA 1996 F0 (Baseline) Focus Condition

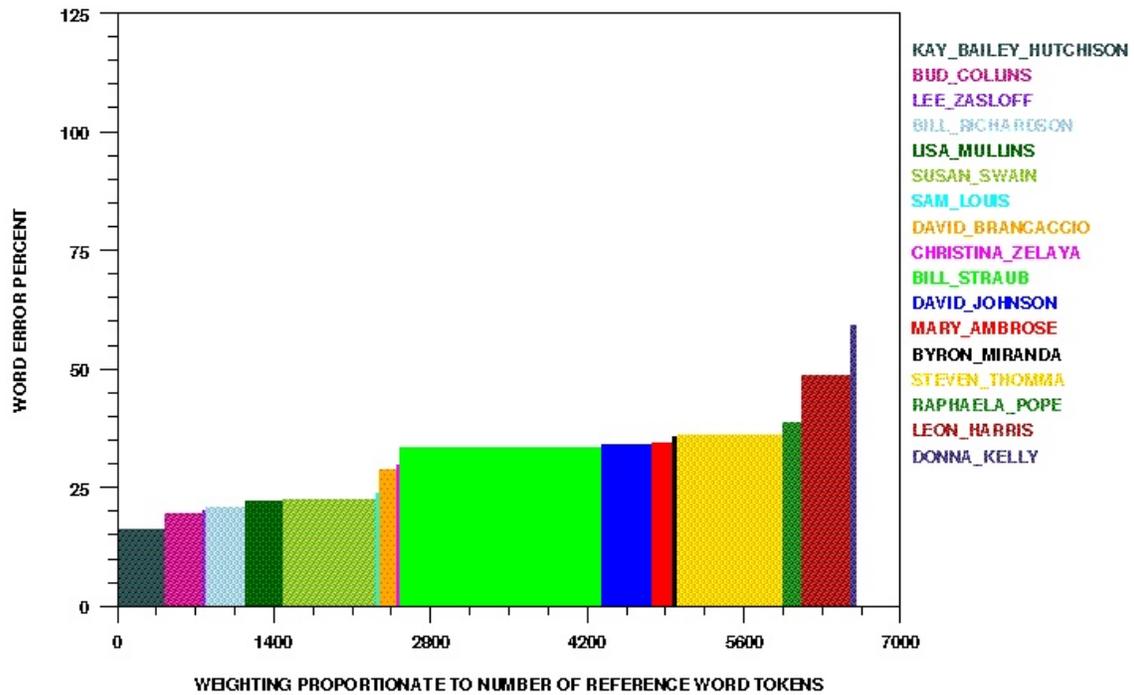


Figure 7(b): IBM1-DARPA 1996 F1 (Spontaneous) Focus Condition

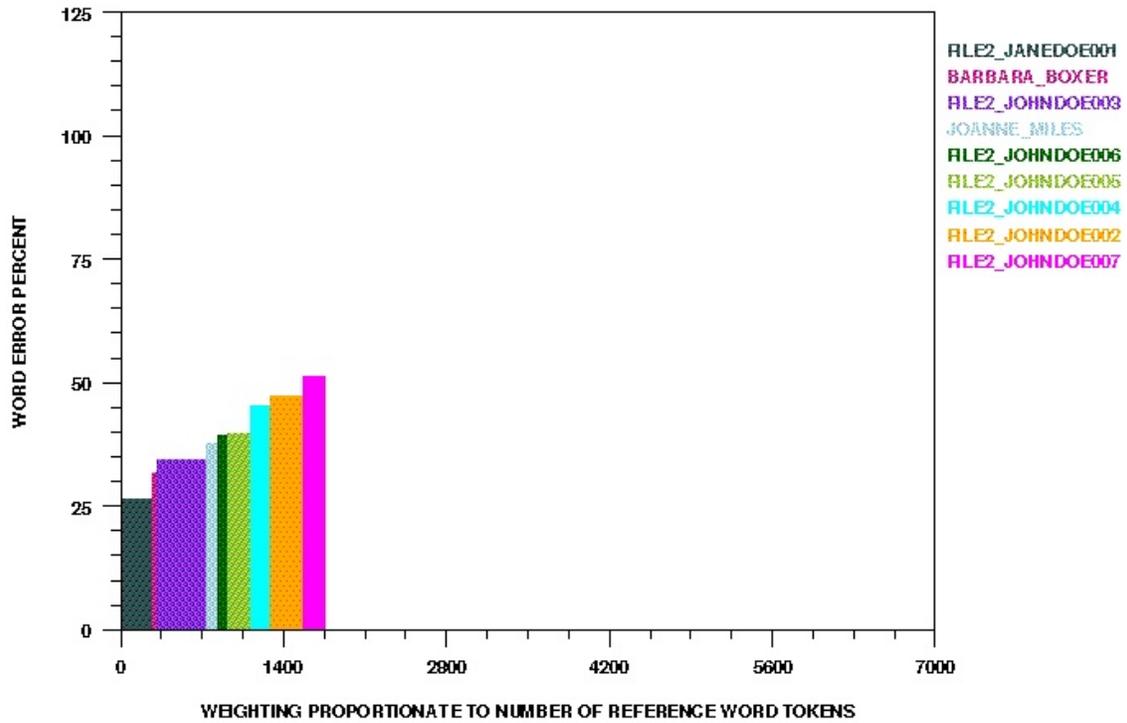


Figure 7(c): IBM1-DARPA 1996 F2 (Telephone) Focus Condition

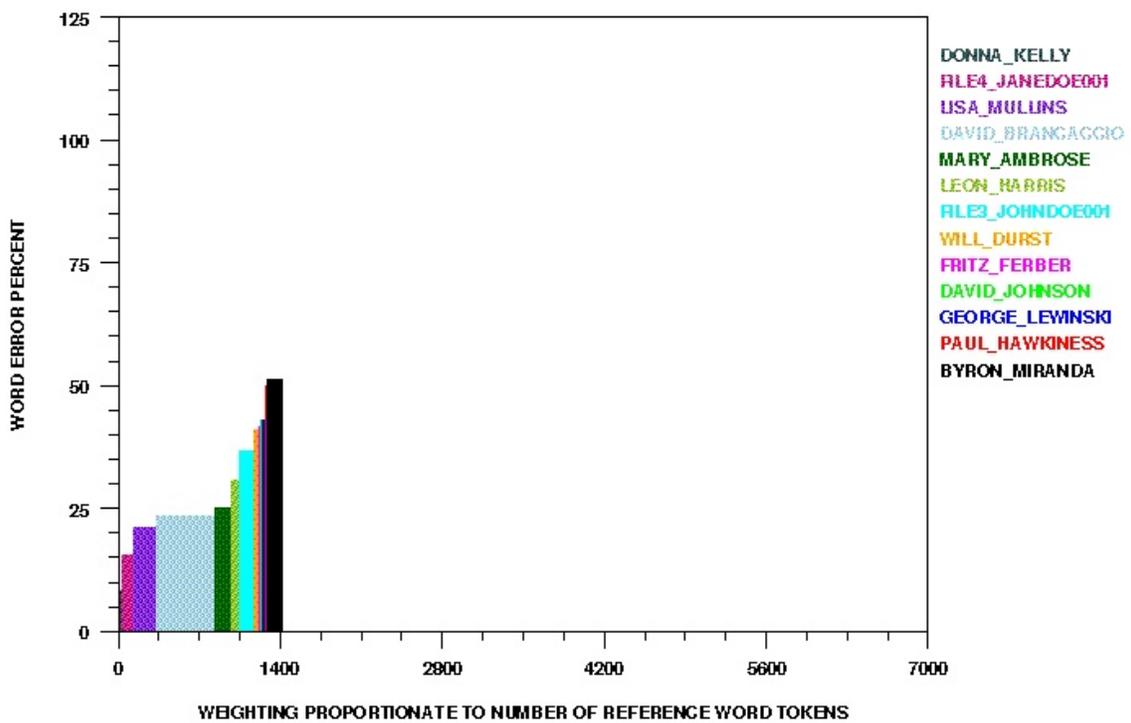


Figure 7(d): IBM1-DARPA 1996 F3 (Bkgrd Music) Focus Condition

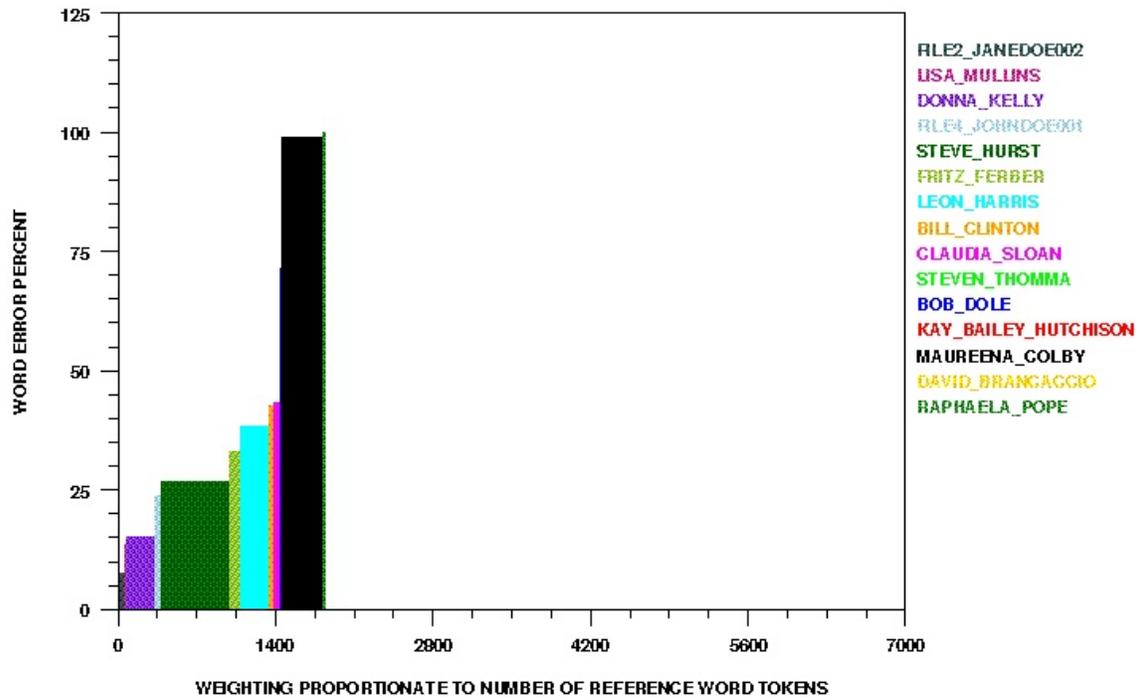


Figure 7(e): IBM1-DARPA 1996 F4 (Bkgrd Noise) Focus Condition

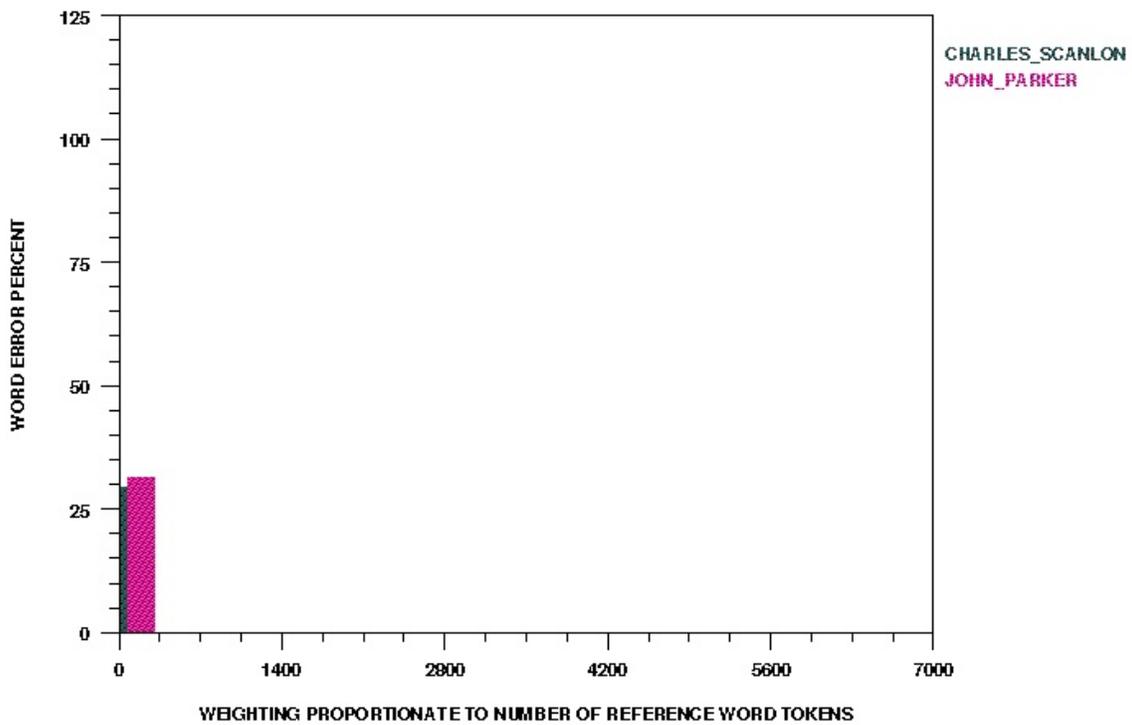


Figure 7(f): IBM1-DARPA 1996 F5 (Nonnative) Focus Condition

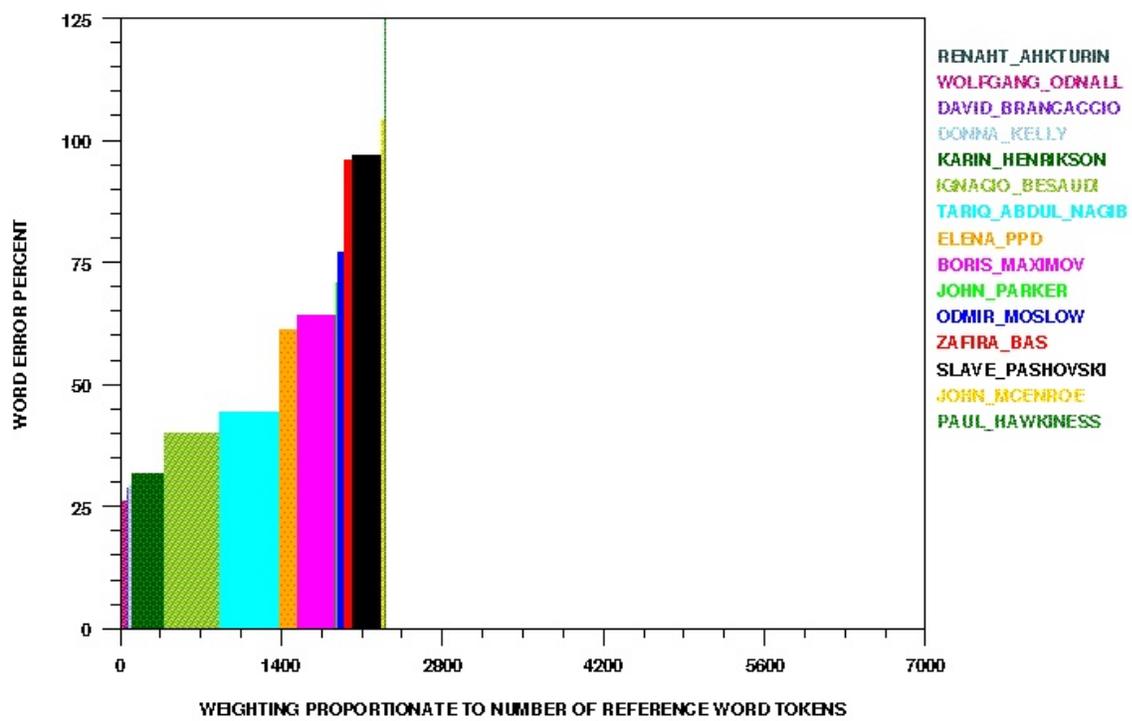


Figure 7(g) IBM1-DARPA 1996 FX (Combination) Focus Condition

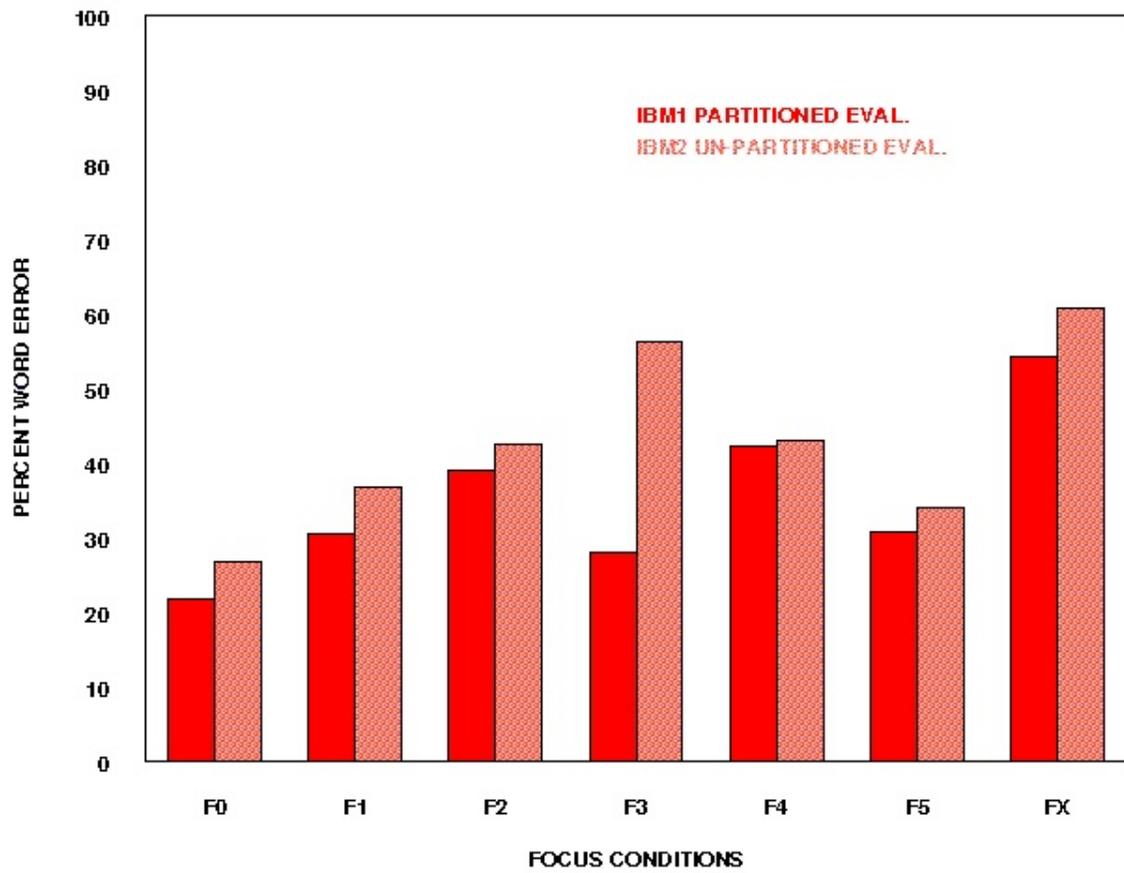


Figure 9: IBM1 PE vs. UE