1998 BROADCAST NEWS BENCHMARK TEST RESULTS: ENGLISH AND NON-ENGLISH WORD ERROR RATE PERFORMANCE MEASURES

David S. Pallett, Jonathan G. Fiscus, John S. Garofolo, Alvin Martin, and Mark Przybocki

National Institute of Standards and Technology (NIST)
Information Technology Laboratory (ITL)
Room A216 Building 225 (Technology)
Gaithersburg, MD 20899
E-mail: dpallett@nist.gov

ABSTRACT

This paper documents the use of Broadcast News test materials in DARPA-sponsored Automatic Speech Recognition (ASR) Benchmark Tests conducted late in 1998.

As in last year's tests [1], statistical selection procedures were used in selecting test materials. Two test epochs were used, each yielding (nominally) one and one-half hours of test material. One of the test sets was drawn from the same test epoch as was used for last year's tests, and the other was drawn from a more recent period.

Results are reported for two types of systems: one (the "Hub", or "baseline" systems) for which there were no limits on computational resources, and another (the "less than 10X real-time spoke" systems) for systems that ran in less than 10 times real-time.

The lowest word error rate reported this year for the "Hub" systems was 13.5%, contrasting with last year's lowest word error rate of 16.2%. For the "less than 10X real-time spoke" systems, the lowest reported word error rate was 16.1%.

Results are also reported, for the second year, on non-English language Broadcast News materials in Spanish and Mandarin.

1. TEST MATERIALS

1.1. English Language Materials

This year's Hub-4E English test set is comprised of two test (sub)sets. Each was selected so as to provide opportunities for year-to-year comparisons of system performance, using "statistically equivalent" test sets. Set 1 was selected from the last year's test pool. The recording dates for Set 1 span from 15 October, 1996 to 14 November, 1996. Set 2 was selected from a 10 hour test pool of broadcast news whose recording dates include June of 1998.

In general, the test materials were chosen using selection criteria documented in Fisher, et al. [2]. As noted in that paper, NIST's efforts toward "balancing of the test pool" from which a random selection was to be made were based on preliminary annotations of the test data by one annotator. In a subsequent reconciliation process that was intended to correct the annotations, the distributions changed, with the result being that the 1997 test set included a larger than expected fraction (larger than in the training material) of the "baseline" (F0) and "spontaneous" (F1) condition speech. This had the effect that the 1997 test set was arguably or unexpectedly "too easy".

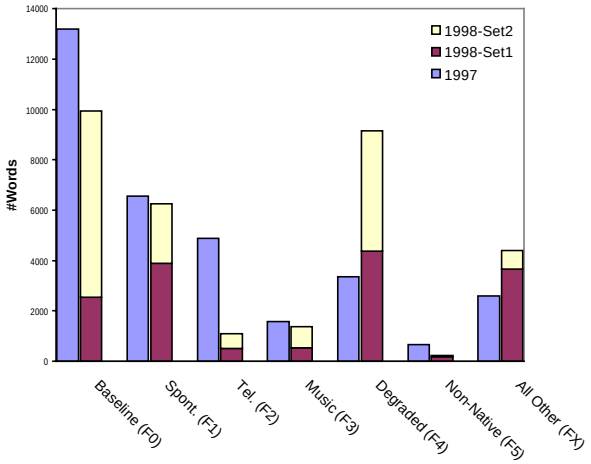

Figure 1. Relative distribution across focus conditions for the 1997 and 1998 test sets.

Figure 1 shows the relative distribution of the 1997 and 1998 test sets for the several focus conditions identified in previous years. For the 1998 test materials (in comparison with the 1997 test set materials) note the existence of: (1) slightly lesser amount of material in the baseline focus condition, (2) lesser proportion of materials in the telephone channel conditions, and (3) the substantially greater proportion of material in degraded acoustics condition, and (4) somewhat greater proportion of materials in the "all other" condition.

Discussion of the relative "difficulty" of the 1997 and 1998 test sets is presented in another section of this paper.

1.2. Non-English Language Material

The test material was drawn from a set of potential test materials provided by the Linguistic Data Consortium that included the following sources:

For the Spanish language, three sources were available: (1) VOA Programming – four original news programs a day, five days a week, (2) ECO – Mexican news show with two reporters in the studio, broadcast on the Galavision network, and (3) Noticiero Univision – half hour weekday news program originating in Miami.

For Mandarin language materials, another three sources were available: (1) VOA Programming – five main programs plus 5-10 minute news slots, (2) CCTV International – evening news broadcast from Bejing, dominated by anchor reading news, and (3) KAZN 1030 AM - all news Los Angeles based Mandarin station.

For each language, selection of test data followed the precedent established last year, involving random selection of stories from a potential test pool and smoothing the transition between stories.

2. EVALUATION PLAN CHANGES

2.1. Evaluation Design Changes

Two changes are notable: (1) the evaluation material specification has been changed to exclude "whole shows", and (2) note that ~200 hours of acoustic training materials are now available from the LDC, vs. last year's ~100 hours. Note also that the same scoring algorithm (SCLITE) used in the 1997 Hub-4 evaluation was used for both the Hub and for the less than 10X spoke.

2.2. "Less than 10X real-time" Spoke

New this year was a spoke involving a challenge to develop more computationally efficient speech recognition algorithms: "systems that run in less than or equal to 10X real-time on a single processor (i.e., less than or equal to ~30 hours to process the ~3 hour evaluation test set)". In the accompanying system description, "system developers [were required to] document all computational resources used for the system, including processor type(s) and memory resources, and including discussion of processing time-allocation for the various signal-processing, segmentation, and decoding components of the system."

The challenge to develop faster systems was motivated by the realization that computational efficiency is important in building successful applications, and that the development of computationally efficient speech recognition algorithms offers genuine technical challenges in its own right. Note that for the baseline systems, run times ranged from ~40 times real-time to as much as ~2000 times real-time, running on machines ranging from 170 MHz Sparc Ultra 1 to as fast as 320 Mips RS6000 systems.

The systems descriptions submitted for these less than 10X real-time systems indicate that, in most cases, the run times were nearly (in most cases, just less than) 10 times real-time, running on (typically) a Pentium II 450 MHz processor, with 512 MB RAM, running either Linux Redhat 4.1 or Windows NT operating systems. One system (CUHTK-Entropic) distributed processing over three processors, two of which were Pentium IIs, and the third a SunUltraSparc II, although total processing times were within the "less than 10 times real-time" limit.

2.3 Information Extraction ("Named Entity") Spoke

A new spoke was added to Hub-4 to examine the effectiveness of broadcast news recognition technology in generating information rich entities and to begin to move the research focus from simple transcription toward spoken information understanding. These entities had been identified by the Message Understanding Conference (MUC) Community as being important for Natural Language and Information Retrieval applications where information is to be extracted from a news stream [3]. The MUC community had worked for several years with entity identification in newswire text and in 1997, a pilot experiment with recognized broadcast news was conducted by MITRE and evaluated with a prototype scoring pipeline, MSCORE which was also developed by MITRE) [4].

Following the MITRE experiment, it was decided that the creation of a common entity tagging task using broadcast news would speed the development of speech recognition technology and include MUC community involvement in developing information extraction technologies for speech applications. Given that the target task was to develop tagging technology for broadcast news, NIST chose to add the task as a spoke to its Hub-4 evaluation to capitalize on the existing infrastructure, corpora, and participant pool. NIST collaborated with MITRE and SAIC to develop the evaluation specifications, corpora, and software. The new task ultimately also required the creation of a new transcription/annotation format for broadcast news. The new spoke was named "Hub-4 Information Extraction - Named Entity" (Hub-4 IE-NE). MITRE and SAIC developed detailed guidelines for the task (Hub-4 Named Entity Task Definition). NIST worked with SAIC to develop scoring software for the task which involved the creation of a Recognition and Extraction Evaluation Pipeline (REEP) to combine the NIST transcription filtering and SCLITE scoring software with the MUC Scorer [5]. The test material was made identical to that for the core tests.

The task involved the recognition and identification of the following types of information entities in the broadcast news stream:

- Named Entities: person, location, organization
- Temporal Expressions: date, time

- Numeric Expressions: monetary, percentage

The Hub-4 IE-NE evaluation included 3 participation levels:

- Full IE-NE: Participants implemented both recognition and entity tagging
- Quasi IE-NE: Participants implemented only entity tagging
- Baseline IE-NE: Participants implemented only recognition

Each participation level specified combinations of the following recognizers and taggers to be evaluated:

- Recognizers: Human reference, CMU SPHINX-III baseline, site recognizer
- Taggers: Human reference, BBN Identifinder , site tagger

In all, six possible recognizer/tagger combinations were evaluated. The participation level and combination approach encouraged wider participation from sites with varying levels of expertise in either recognition or entity tagging and it permitted NIST to evaluate the recognition and entity tagging components separately.

Further details including the development of the IE-NE spoke and the scoring and analysis of the results of the evaluation are given in [5].

3. PARTICIPANTS

There were nine research sites participating in the traditional Broadcast News Hub transcription task: GTE Internetworking's BBN Technologies, Cambridge University's Engineering Department HTK group (CU-HTK), Dragon Systems (DRAGON), IBM's T.J. Watson Laboratories (IBM), the French National Laboratories' Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), a ollaborative effort involving the Oregon Graduate Institute and fonix Corporation (OGI_FONIX), a joint effort involving Philips Research Laboratories Aachen and Lehrstuhl fuer Informatik VI Rheinisch-Westfaelische Technische Hochschule Aachen (PHILIPS_RWTH), a European Union funded project entitled "Speech Recognition Algorithms for Connectionist Hybrids" involving Cambridge University's Engineering Department, Sheffield University, and the International Computer Science Institute (SPRACH), and SRI International (SRI).

The six participants in the "less than 10 times real-time" spoke included: BBN, a collaborative effort involving Cambridge University's HTK group and Entropic Ltd. (CUHTK-Entropic), DRAGON, IBM, SPRACH and SRI.

There were four participants in the non-English language tests: BBN, CMU, Dragon Systems, and IBM. BBN and CMU participated in the Spanish tests, and Dragon and IBM participated in the Mandarin tests.

4. TEST RESULTS

4.1. Automatic Transcription Hub

The test plan states that "Special attention will be given to the F0 condition. This condition is of particular interest because the absence of other complicating factors such as background noise, music and non-native dialects focuses attention on basic speech recognition issues common to all conditions." The F1 focus condition is also of interest because it also lacks complicating factors such as noise, music and non-native dialects, but includes evidence of spontaneity such as disfluencies.

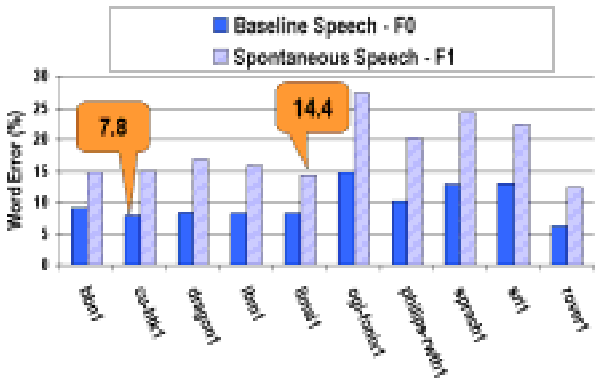

Figure 2. Word error rates for the low noise baseline and spontaneous spokes.

Figure 2 shows the word error rates reported by the developers of the Hub systems for the low-noise baseline, F0, and spontaneous, F1, conditions. The lowest word error rate for the baseline speech was 7.8%, reported for the CU-HTK system. The LIMSI system achieved the lowest word error rate for the spontaneous speech, 14.4%.

The test plans also state that "NIST will tabulate and report word error rates over the entire dataset."

Figure 3 shows the results of a rank-ordering of word error rate results for the entire 1998 dataset for the Hub systems (including the NIST-implemented ROVER results). Results are shown for both of the test sets comprising the 1998 test set as well as for the overall test set word error rate. Ovals are used to indicate that differences in reported word error rates are not shown to be significant, using the NIST MAtched Pair Sentence Segment Word (MAPSSWE) Error Paired Comparison Significance test. For example, differences in word error rate are not shown to be significant for the IBM, LIMSI, and CU-HTK systems. Performance differences between Dragon Systems and BBN are not shown to be significant, as is also the case for SPRACH and SRI.

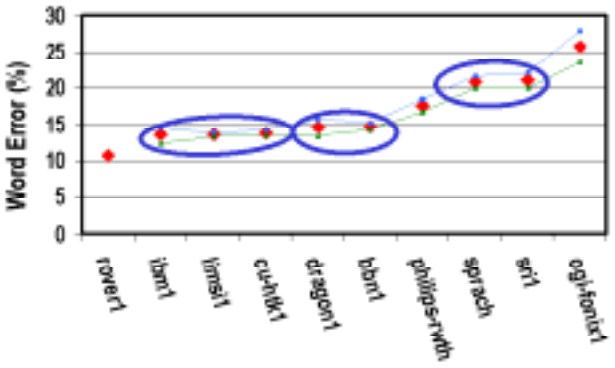

Figure 3. Systems ordered by overall word error rate.

Table 1 documents the error rates found for the "Hub" systems, with word error rates ranging from 13.5% for the IBM1 system to 25.7% for the OGI_fonix system.

Note that this table includes word error rate found for each focus condition in addition to the overall word error rate.

Table 2 provides a tabulation of the several significance tests that are implemented by NIST, in this case, for the Hub systems using the overall word error rate in the comparisons.

4.2. "Less than 10X real-time" Spoke

For the "less than 10 times real-time" systems, Figure 4 shows the reported word error rates for the six "less than 10 times real-time" systems, for the low-noise baseline and spontaneous conditions. The lowest word error rate for the baseline speech was 9.7%, achieved by the CUHTK-Entropic system, and for the spontaneous speech it was 17.0%, achieved by the BBN system.

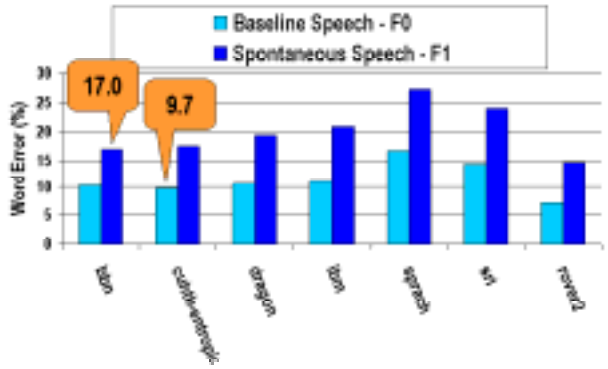

Figure 4. Word error rates for the low noise baseline and spontaneous spokes for the less than 10X real-time systems.

Figure 5 shows the results of a rank-ordering of results for the "less than 10 times real-time" systems by word error rate. As in Figure 2, an oval is used to indicate that differences in reported word error rates are not shown to be significant, using the MAPSSWE test. In this case, performance differences between the Dragon and BBN systems are not significant.

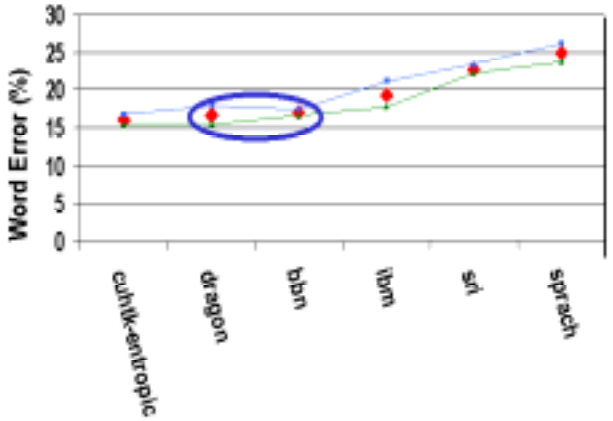

Figure 5. Less than 10X real-time systems ordered by overall error rate.

Table 3 indicates the results reported for the "less than 10 times real-time" systems. Word error rates range from a minimum for 16.1% for the CUHTK-Entropic system, to 25.0% for the SPRACH2_10X system.

Table 4 provides a tabulation of the several significance tests that are implemented by NIST, in this case for the less than 10X real-time systems, and the overall word error rate.

4.3. Non-English Transcription Task

Spanish

The word error rate reported for the BBN Technologies system for the Spanish language test set was 21.5%, contrasting with 20.3% for last year's test set. The word error rate for this year's CMU system was 22.4%, in contrast with last year's error rate of 23.5%.

Mandarin

The character error rate reported for the Dragon Systems' Mandarin system was 20.6%, which contrasts with last year's error rate of 20.2%. The character error rate reported for this year's IBM Mandarin system was 17.1%, in contrast with last year's test results of 19.8%.

5. DISCUSSION

5.1. Differences Between 1997 and 1998 Test Sets

Recall that when comparing the relative amounts of material in the various focus conditions for the 1997 and 1998 test sets, there was markedly less "telephone channel" material in the 1998 test set, and markedly more in the "degraded acoustics" focus condition. The first of the comparisons suggests the 1998 test set would be easier than the 1997 test set, and the second of these comparisons suggests that the 1998 test set would be

harder. Thus a comparison of the relative difficulty of two test sets might best be made with the use of the same reference algorithm, operating on the two test sets in question.

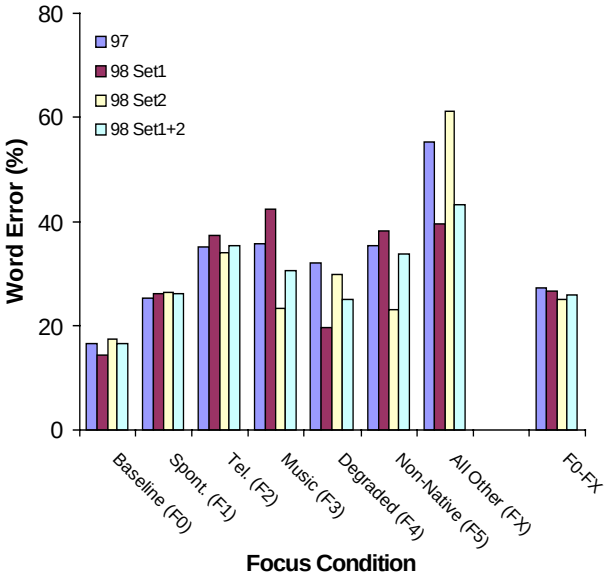

Figure 6. Error rates for the 1997 and 1998 test sets (CMU-developed Sphinx III recognizer).

NIST has a copy of the CMU-developed Sphinx III Broadcast News System, and processed both the 1997 and 1998 test sets with this system. Figure 6 shows the error rates for both the 1997 and 1998 test sets (along with error rates found for the two subsets of the 1998 test set).

Focusing attention on the low-background noise F0 condition, the word error rate for the 1997 test set was 16.7% and in 1998, it was also 16.7%. In the F1 condition, the 1997 error rate was 25.4%, and in 1998, it was 26.2%. The overall word error rate (F0-FX) for the 1997 test set is 27.1%, and for the 1998 test set is 25.8%, suggesting that, over all focus conditions, the 1998 test set is slightly easier than the 1997 set.

These comparisons suggest that the two test sets (the 1997 and 1998 test sets) are very comparable, although not identical, in difficulty.

5.2. Implementations of ROVER

The NIST-developed software system for combining alternative transcriptions [6] was implemented at five of the nine "core" systems: (1) BBN's core system implemented four decodings (with different frame rates) and combined them with ROVER, (2) CU-HTK's core system annotated lattices and 1-best outputs with confidence scores and combined them with ROVER, (3) Dragon Systems' core system ran two different types of recognizer, differing in the type of recognizer that was used in the chopping step in the beginning (one with standard triphone recognizer, and the other used left diphone models without cross-word co-articulation) and the outputs were combined with ROVER, (4) IBM's core system merged seven hypothesized scripts (involving several forms of adaptation and four baseline systems) using ROVER, and (5) SPRACH's core system produced hypotheses from three acoustic models (2 context independent, and one involving 676 word-internal context-dependent phone probabilities) and these hypotheses were merged with ROVER.

Of the less than 10 times real-time systems, only the SPRACH system implemented the ROVER software.

At NIST, using submitted results files, ROVER was used to generate two combined systems hypothesis files – one using the "core" Hub systems results, and another using the less than 10 times real-time systems results. As shown in Figure 3, the word error rate for the ROVER implementation for the Hub systems results was 10.6%.

ACKNOWLEDGEMENTS

We would like to acknowledge the assistance of Audrey Le and Bill Fisher, in selecting and screening the English-language test materials and checking the transcriptions and annotations. Special thanks are also due to Alberto Arroyo and Mei Alsop who verified, corrected and annotated the transcriptions for the Spanish and Mandarin test materials.

NOTICE

The views expressed in this paper are those of the authors. The test 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 test materials, selecting and defining reference annotation and transcriptions files for the tests, developing and implementing scoring software, and uniformly scoring and tabulating results. The views of the authors, 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., et al., "1997 Broadcast News Benchmark Test Results: English and Non-English," *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 8-11, 1998, Lansdowne VA, pp. 5 – 11.

[2] Fisher, W., et al., "Data Selection for Broadcast News CSR Evaluations," *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 8-11, 1998, Lansdowne VA, pp. 12 – 15.

[3] Chinchor, N., "Overview of MUC-7 Proc.," *Message Understanding Conference 7*, 1998.

[4] Burger, J., Palmer, D., Hirschman, L., "Named Entity Scoring for Speech Input," *Proc. 36'th Annual Meeting of the Association for Computational Linguistics (ACL/COLING '98)*, August 1998.

[5] Przybocki, M., Fiscus, J., Garofolo, J., Pallett, D.,"1998 Hub-4 Information Extraction - Named Entity

Evaluation," to be appeared in *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 28 - March 3, 1999, Dulles, VA.

[6] Fiscus, J.G., "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara CA, pp. 347-354.

Table 1. Word error rates, overall and for the several focus conditions, for the Hub or baseline systems.

SYSTEM	Overall	Baseline Broadcast Speech	Spontaneous Broadcast Speech	Speech Over Telephone Channels	Speech in the Presence of Background Music	Speech Under Degraded Acoustics Conditions	Speech from Non-Native Speakers	All Other Speech	Female	Male
	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE
	Hub4 Focus Conditions								Speaker Sex	
	Set/Subset #Words and System Set/Subset Average Word Error Rate									
bbn1_at.ctm	[32443] 14.7	[9948] 9.0	[6247] 15.0	[1095] 20.6	[1385] 19.2	[9145] 13.9	[235] 17.9	[4388] 25.9	[13165] 14.1	[19250] 14.7
cu-htk1_at.ctm	[32443] 13.8	[9948] 7.8	[6247] 15.1	[1095] 20.1	[1385] 15.8	[9145] 13.6	[235] 16.6	[4388] 24.1	[13165] 12.5	[19250] 14.3
dragon1_at.ctm	[32443] 14.5	[9948] 8.3	[6247] 16.8	[1095] 19.0	[1385] 15.2	[9145] 13.4	[235] 24.3	[4388] 25.6	[13165] 13.9	[19250] 14.4
ibm1_at.ctm	[32443] 13.5	[9948] 8.2	[6247] 16.0	[1095] 17.4	[1385] 17.3	[9145] 12.1	[235] 15.3	[4388] 22.1	[13165] 13.5	[19250] 12.9
limsi1_at.ctm	[32443] 13.6	[9948] 8.2	[6247] 14.4	[1095] 16.9	[1385] 16.3	[9145] 13.6	[235] 21.3	[4388] 22.2	[13165] 12.5	[19250] 13.8
ogi_fonix1_at.ctm	[32443] 25.7	[9948] 14.9	[6247] 27.3	[1095] 38.3	[1385] 33.4	[9145] 24.8	[235] 29.4	[4388] 44.0	[13165] 26.4	[19250] 24.6
philips_rwth1_at.ctm	[32443] 17.6	[9948] 10.1	[6247] 20.2	[1095] 25.6	[1385] 22.1	[9145] 16.4	[235] 29.4	[4388] 29.5	[13165] 16.8	[19250] 17.7
sprach1_at.ctm	[32443] 20.8	[9948] 13.1	[6247] 24.3	[1095] 30.2	[1385] 24.5	[9145] 19.4	[235] 24.3	[4388] 32.7	[13165] 20.6	[19250] 20.4
sri1_at.ctm	[32443] 21.1	[9948] 13.2	[6247] 22.4	[1095] 25.9	[1385] 23.3	[9145] 20.5	[235] 25.5	[4388] 36.0	[13165] 20.3	[19250] 20.9

Table 2. Tabulation of the several significance tests. Hub or baseline systems.

Composite Report of All Significance Tests For the 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 (Sentence Error)	MN

Test Abbrev.		bbn1_at	cu-htk1_at	dragon1_at	ibm1_at	limsi1_at	ogi_fonix1_at	philips_rwth1_at	sprach1_at	sri1_at	Test Abbrev.
MP	bbn1_at		cu-htk1_at	~	ibm1_at	limsi1_at	bbn1_at	bbn1_at	bbn1_at	bbn1_at	MP
SI			~	~	ibm1_at	~	bbn1_at	bbn1_at	bbn1_at	bbn1_at	SI
WI			~	~	ibm1_at	limsi1_at	bbn1_at	bbn1_at	bbn1_at	bbn1_at	WI
MN			~	~	~	~	bbn1_at	bbn1_at	bbn1_at	bbn1_at	MN
MP	cu-htk1_at			cu-htk1_at	~	~	cu-htk1_at	cu-htk1_at	cu-htk1_at	cu-htk1_at	MP
SI				~	ibm1_at	~	cu-htk1_at	cu-htk1_at	cu-htk1_at	cu-htk1_at	SI
WI				~	ibm1_at	~	cu-htk1_at	cu-htk1_at	cu-htk1_at	cu-htk1_at	WI
MN				~	~	~	cu-htk1_at	cu-htk1_at	cu-htk1_at	cu-htk1_at	MN
MP	dragon1_at				ibm1_at	limsi1_at	dragon1_at	dragon1_at	dragon1_at	dragon1_at	MP
SI					ibm1_at	limsi1_at	dragon1_at	dragon1_at	dragon1_at	dragon1_at	SI
WI					ibm1_at	limsi1_at	dragon1_at	dragon1_at	dragon1_at	dragon1_at	WI
MN					~	~	dragon1_at	~	dragon1_at	dragon1_at	MN
MP	ibm1_at					~	ibm1_at	ibm1_at	ibm1_at	ibm1_at	MP
SI						~	ibm1_at	ibm1_at	ibm1_at	ibm1_at	SI
WI						~	ibm1_at	ibm1_at	ibm1_at	ibm1_at	WI
MN						~	ibm1_at	~	ibm1_at	ibm1_at	MN
MP	limsi1_at						limsi1_at	limsi1_at	limsi1_at	limsi1_at	MP
SI							limsi1_at	limsi1_at	limsi1_at	limsi1_at	SI
WI							limsi1_at	limsi1_at	limsi1_at	limsi1_at	WI
MN							limsi1_at	limsi1_at	limsi1_at	limsi1_at	MN
MP	ogi_fonix1_at							philips_rwth1_at	sprach1_at	sri1_at	MP
SI								philips_rwth1_at	sprach1_at	sri1_at	SI
WI								philips_rwth1_at	sprach1_at	sri1_at	WI
MN								philips_rwth1_at	~	~	MN
MP	philips_rwth1_at								philips_rwth1_at	philips_rwth1_at	MP
SI									philips_rwth1_at	philips_rwth1_at	SI
WI									philips_rwth1_at	philips_rwth1_at	WI
MN									philips_rwth1_at	philips_rwth1_at	MN
MP	sprach1_at									~	MP
SI										~	SI
WI										~	WI
MN										~	MN
MP	sri1_at										MP
SI											SI
WI											WI
MN											MN

Table 3. Word error rates, overall and for the several focus conditions, for the less than 10X real-time systems.

SYSTEM	Overall	Hub4 Focus Conditions: Baseline Broadcast Speech	Spontaneous Broadcast Speech	Speech Over Telephone Channels	Speech in the Presence of Background Music	Speech Under Degraded Acoustics Conditions	Speech from Non-Native Speakers	All Other Speech	Speaker Sex: Female	Speaker Sex: Male
	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE	#Wrd %WE
	Set/Subset #Words and System Set/Subset Average Word Error Rate									
bbn2_10x.ctm	[32443] 17.1	[9948] 10.3	[6247] 17.0	[1095] 24.9	[1385] 22.5	[9145] 16.5	[235] 21.7	[4388] 29.7	[13165] 17.2	[19250] 16.5
dragon2_10x.ctm	[32443] 16.7	[9948] 10.6	[6247] 19.5	[1095] 23.6	[1385] 21.2	[9145] 14.4	[235] 25.5	[4388] 27.9	[13165] 16.0	[19250] 16.6
cuhtk-entropic1_10x.ctm	[32443] 16.1	[9948] 9.7	[6247] 17.6	[1095] 19.1	[1385] 19.5	[9145] 15.7	[235] 23.4	[4388] 27.3	[13165] 15.0	[19250] 16.3
ibm4_10x.ctm	[32443] 19.4	[9948] 11.0	[6247] 20.9	[1095] 28.8	[1385] 25.1	[9145] 18.0	[235] 23.0	[4388] 35.2	[13165] 20.5	[19250] 17.8
sprach2_10x.ctm	[32443] 25.0	[9948] 16.8	[6247] 27.3	[1095] 35.5	[1385] 33.4	[9145] 22.7	[235] 32.8	[4388] 39.2	[13165] 25.2	[19250] 23.9
sri2_10x.ctm	[32443] 22.8	[9948] 14.4	[6247] 24.1	[1095] 28.4	[1385] 25.7	[9145] 22.9	[235] 27.2	[4388] 36.9	[13165] 22.0	[19250] 22.7

Table 4. Tabulation of the several significance tests. Less than 10X real-time systems.

Composite Report of All Significance Tests
For the DARPA CSR 1998 Test Sets 1 and 2, Less Than 10X Primary Systems 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 (Sentence Error)	MN

Test Abbrev.		bbn2_10x	dragon2_10x	cuhtk-entropic1_10x	ibm4_10x	sprach2_10x	sri2_10x	Test Abbrev.
MP	bbn2_10x		~	cuhtk-entropic1_10x	bbn2_10x	bbn2_10x	bbn2_10x	MP
SI			~	~	bbn2_10x	bbn2_10x	bbn2_10x	SI
WI			~	cuhtk-entropic1_10x	bbn2_10x	bbn2_10x	bbn2_10x	WI
MN			~	cuhtk-entropic1_10x	~	bbn2_10x	bbn2_10x	MN
MP	dragon2_10x			cuhtk-entropic1_10x	dragon2_10x	dragon2_10x	dragon2_10x	MP
SI				~	dragon2_10x	dragon2_10x	dragon2_10x	SI
WI				~	dragon2_10x	dragon2_10x	dragon2_10x	WI
MN				cuhtk-entropic1_10x	~	dragon2_10x	dragon2_10x	MN
MP	cuhtk-entropic1_10x				cuhtk-entropic1_10x	cuhtk-entropic1_10x	cuhtk-entropic1_10x	MP
SI					cuhtk-entropic1_10x	cuhtk-entropic1_10x	cuhtk-entropic1_10x	SI
WI					cuhtk-entropic1_10x	cuhtk-entropic1_10x	cuhtk-entropic1_10x	WI
MN					cuhtk-entropic1_10x	cuhtk-entropic1_10x	cuhtk-entropic1_10x	MN
MP	ibm4_10x					ibm4_10x	ibm4_10x	MP
SI						ibm4_10x	ibm4_10x	SI
WI						ibm4_10x	ibm4_10x	WI
MN						ibm4_10x	~	MN
MP	sprach2_10x						sri2_10x	MP
SI							sri2_10x	SI
WI							sri2_10x	WI
MN							sri2_10x	MN
MP	sri2_10x							MP
SI								SI
WI								WI
MN								MN