

W.S. Pallatt

1995 HUB-4 "DRY RUN" BROADCAST MATERIALS BENCHMARK TESTS

David S. Pallatt, Jonathan G. Fiscus, John S. Garofolo, Mark A. Przybocki

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

ABSTRACT

This paper documents first usage of broadcast materials in ARPA-sponsored Automatic Speech Recognition (ASR) benchmark tests. The materials used for these initial "dry run" tests were derived from "Marketplace" radio broadcasts. ("Marketplace" is produced at radio station KUSC in Los Angeles, and distributed by Public Radio International and deals with business news. Information about "Marketplace" can currently be found on the World Wide Web at "<http://www.usc.edu/marketplace>".) With the assistance of the Linguistic data Consortium and KUSC, NIST prepared a limited amount of Marketplace-derived materials for the use of researchers and implemented test protocols agreed to by four participating sites. This paper documents a number of complementary measures of error for various test sets and subsets, and illustrates some properties of the measured word errors for one of the well-performing systems.

1. INTRODUCTION

For several years, much of the research within the ARPA-sponsored automatic speech recognition community has focussed on large-vocabulary continuous "read" (vs. "spontaneous") speech, with use of language models largely derived from text corpora dealing with business-oriented news (i.e., "Wall Street Journal-based", or, more generally, North American Business news). Only a very limited amount of material representing "spontaneous dictation" has been available and studied. Furthermore, the bulk of the research has made use of acoustic training materials which, like the test materials, also consist of "read speech", with prompting texts derived from related text corpora.

To this time, little attention has been paid to automatic recognition of the type of speech found in radio or television broadcasts, even for broadcasts dealing with business news. The speech in these broadcasts includes a wide variety of speaking styles (e.g., read news texts from "anchors", studio announcers, or from correspondents and in interviews), dialects (both regional and foreign-accented English), and exhibits background noise and channel effects (including the presence of background music and reduced bandwidth -- apparently telephone -- speech). These broadcasts (like the text corpora, but perhaps to a greater degree) may not be strictly limited to "business-oriented" news..

Early in 1995, researchers at NIST recorded an "off-the-air" radio broadcast of one "Marketplace" program.. NIST staff selected several segments from the broadcast and down-sampled the off-the-air 48 kHz sample rate Digital Audio Tape (DAT) recording to 16 kHz, and then excised and transcribed sentence-length files. These files were then processed using a (not-quite-state-of-the-art)

CMU-developed Sphinx-II HMM recognizer that had been trained for use with North American Business News texts, and which used a 20K word NAB news-derived trigram language model. We observed interesting phenomena: there was marked variability in performance across different speakers and the different speech styles, and in general, error rates were substantially higher than for "read" speech. However, for some subjects, performance seemed comparable to that observed with some of the "read" speech used for previous tests. The variability in error rate for these broadcasts, and finding the causes of that variability, would appear to be important to study.

The use of broadcast materials for ARPA-community Benchmark Tests in November 1995, specifically materials derived from "Marketplace" broadcasts, was discussed last May at ICASSP-95 in Detroit. Government research sponsors indicated interest in these materials and tests -- especially if only a severely limited amount of domain-specific acoustic training material and transcriptions of Marketplace broadcasts was required by the system developers. Subsequently, at ARPA's request, the Linguistic Data Consortium made arrangements for NIST to receive DAT copies of one year's "Marketplace" broadcasts, and for a transcription service (Cambridge Transcripts) to work with NIST staff to transcribe a subset of those broadcasts. In mid-May, a 16 kHz sample rate file derived from the broadcast of November 12, 1993 was distributed by NIST on CD-ROM to twelve research sites that had indicated tentative interest in this challenge.

Concurrently, a working group was formed, involving representatives of several sites, chaired by Alex Rudnicky at CMU, to develop the specifications for the November 1995 Benchmark Tests. It was agreed that these tests were to be referred to as "dry run" tests since much of what was to be tried was without precedent, and only limited resources were available for this research.

By mid-Summer, 1995, the following text had been agreed upon to describe the purpose of these tests:

"The purpose of Hub-4 is to encourage research into those aspects of speech recognition technology that make for nimbleness, the ability of systems to adapt to varying conditions of input, whether in acoustic characteristics or content. Equally, Hub-4 will focus on the problems of processing "found speech", that is, speech materials which have not been created specifically for the purpose of speech system development or evaluation. In other words, Hub-4 is meant to promote research on

adaptation to changing conditions and on robustness with respect to degradation.”

This paper is intended to document the preparation of the intentionally limited amount of Marketplace-broadcast-domain-specific training and test materials, and to tabulate some of the results obtained by the four sites that participated in these “dry run” tests -- BBN, CMU, Dragon Systems, and IBM. Because of the diverse nature of the “found speech”, and because no general agreement has yet been reached as to the value of any “single number” metric, the reader is encouraged to look at the range of any set of results, rather than any one number, and to note the qualitative nature of the phenomena found in these preliminary studies.

2. HUB-4 MARKETPLACE TRAINING AND TEST MATERIALS

As indicated in the previous section, after an agreement between the LDC and KUSC had been reached, KUSC shipped DAT tapes derived from KUSC’s archives to NIST. At NIST, conventional audio cassette tapes were made, and the DAT tapes were down-sampled to provide 16 kHz SPHERE-headered whole-broadcast files.

The audio cassette tapes were sent to a transcription service (Cambridge Transcripts), along with preliminary specifications for the format of the transcriptions. When the transcriptions were received at NIST, they were checked programmatically for format compliance and corrections were made. For the training data, story-boundary time marks were added and the data was distributed to the research sites. The development test data was processed more intensively. The transcriptions and speaker information files were hand-verified and corrected for format and content. Story-boundary and speaker-boundary time marks were then added to the transcriptions. The evaluation test data was processed similarly to the development test data. Music boundary time marks were also added to the evaluation test data so that music effects could be isolated during scoring. [Representative evaluation transcript and speaker information data is included in Appendices 2 and 3]

Other information included in derivative files used for scoring included annotations about apparently reduced bandwidth (presumably, but not necessarily due to telephone channels), and the presence of foreign-accented English.

A preliminary decision had been made to adopt the “training epoch” cited for the November 1994 NAB News Benchmark tests (predating April 1, 1994) for these studies. In doing so, it was thought that existing language models for this epoch might provide a useful starting point for these studies. It was suggested that the “Marketplace” domain-specific “training” materials be selected from this epoch. Accordingly, NIST staff selected a total of 10 broadcasts from the period predating May 15, 1994, and prepared a CD-ROM that included 16 kHz files derived from the corresponding DATs provided by KUSC. Verified transcriptions were also made available by NIST. These 10 broadcasts, spanning the period between November 17, 1993 and March 28, 1994, comprised the only domain-specific materials designated for

system training.

Concurrently, the Working Group had agreed on a number of things, including preliminary terminology and the composition of a “test set”. It was agreed that a test set was to be comparable in size to approximately three half-hour broadcasts, but in fact consisting of: (a) one complete broadcast, plus (b) another two initial portions, termed the “heads” of the broadcasts, from just after the introductory material and continuing to just prior to the stock market summary (introduced by the phrase “Let’s do the numbers”), plus © another two concluding portions, termed the “tails” of the broadcasts, from just after the stock market summary to the end of the last story, but not including the closing segment and credits. NIST staff prepared two development test sets according to these criteria, selected from the “dev test” epoch for the 1994 tests (15 May - 15 June 1994), and distributed these on CD-ROM on July 24, 1995 to seven potential participating sites.

Government sponsors and advocates of this research had requested that the evaluation test materials be drawn from contemporary sources. Accordingly, it was agreed that the evaluation test materials would be drawn from the period 1 - 31 August, 1995, the “test epoch” for these tests. The LDC arranged for KUSC to provide additional materials spanning this period, and NIST staff then selected two evaluation test sets from the test epoch. One was completely processed and distributed on CD-ROM by NIST on November 1, 1995 to the four sites participating in the November 1995 “Dry Run” tests. Material for the second test set (consisting of one complete broadcast, two “heads” and two “tails”) has been kept in NIST’s “archives” for potential future use, possibly in November 1996.

The 5 Marketplace evaluation test shows were distributed -- in their entirety -- to the participating sites along with an index file which specified the segments to be used in the test. The test set consisted of 71.4 minutes of broadcast data drawn from the 5 shows as follows:

Broadcast	Portion	Beg Time	End Time	Duration
950801	complete	0.00	1725.94	1725.94
950811	head	82.74	671.08	588.34
950814	tail	1038.18	1661.36	623.18
950824	tail	1079.65	1667.47	587.82
950830	head	79.61	839.65	760.04

(times are in seconds)

The 5 complete Marketplace shows contained a total of 144 minutes of broadcast data and are each about 5.5Mb in size. The test set contained 21 stories and speech from 36 unique speakers. The speech contained a total of 12,518 lexical tokens consisting of 2,719 unique lexemes.

3. HUB-4 TEST PARADIGM

Another paper in this Proceedings documents the test protocols that were agreed to by the Hub-4 Working Group, and were approved by government sponsors and advocates of this research.

After the November 1994 benchmark tests, NIST staff made a number of changes to the NIST string alignment and scoring software, principally motivated by concerns expressed by some users that it was unnecessarily complex and included too many rarely invoked options. The revised software is referred to as "sclite", and was used for all of the 1995 CSR benchmark tests, including the Hub-4 tests.

The reference transcripts for the Hub-4 test material had been time-marked with the times corresponding to story, speaker, and music boundaries. System output hypothesis files were provided with time-markings for each hypothesized word. In the string-alignment process implemented by sclite, each segment defined by the boundary time-markings was regarded as an "utterance", regardless of its length. Conventional dynamic-programming procedures were used to align the reference and hypotheses strings corresponding to each turn, without reliance on other time marking or phonologic information. In the vicinity of boundaries, the times corresponding to the mid-points of individual hypothesized words were used to assign these words to one of the reference segments. Scoring, per se, was performed at the word-error level only for each "segment". As will be seen, considerable variability was observed in error rate from one segment to another, even within one "story".

The hypothesis files submitted by the several participants in these tests exhibited a number of lexicographic inconsistencies for which canonical representations could not be determined from the transcription rules, particularly involving compound words (e.g., "buyout" vs. "buy out"), variant spellings (e.g., "ok" vs. "okay"), proper name spelling inconsistencies and variants (e.g., "Cravis" vs. "Kraavis"), and cases where the language model apparently conflicted with the transcription rules (e.g., "in house" vs. "inhouse"). Rather than score these as errors, a prefiltering operation was used to pre-process the hypothesis and reference files before scoring.

As in previous tests, sites were permitted to submit "requests for adjudication" after the preliminary scoring. During adjudication, sites submitted bug reports to NIST via Email. These adjudication bug reports usually request that a transcription be "corrected" or that an alternation be permitted. Sites were also permitted to comment on or contest another site's requests. This process, although laborious, ensured that the transcription of the test data and resultant scored results were as accurate as possible.

Only 1 of the 4 Hub-4 participating sites chose to submit adjudication requests. Of the 96 adjudication requests received, 31 were granted, 2 were partially granted, and 63 were denied. Most of the requests that were denied were requests for forgiving obvious homophone errors. These have always been disallowed in CSR tests unless they are contextually ambiguous, or it is clear that no constraining grammar was to have been used (i.e., homophone errors were allowed in the Resource Management "No Grammar" tests). It should be noted that the adjudication resulted in a

decrease in error rate of only about 0.2% in most systems' results and resulted in no change to the relative ranking of the different systems.

4. HUB-4 TEST RESULTS

Recognizing that the basis for discussion in these "dry run" tests had been agreed to be conventional word error measures, and that each test involves many segments -- each with potentially different speaking styles, the presence or absence of background music to varying degree within segments, and differing channel or bandwidth -- it should be evident that it is inappropriate to summarize results in any one "single number" measure of error.

Measurement of "word error rate" is a generally accepted procedure in benchmark tests.

In many previous tests, care has been taken to balance the amount, and in some cases the nature, of the test material from each speaker in a test set. Reported word error rates are often obtained from simply counting all word errors, and dividing that number by the number of word tokens in the relevant test set. Such an approach may be used for the subsets of materials in an individual's test material (i.e., subsets comprising 15 sentence utterances spoken by each of 20 individual speakers in this year's Hub 3 tests), or for the aggregate set of material in a larger test set (i.e., the set of ~300 utterances in the complete test set for this year's Hub 3 C0 tests).

But when the amount of material from each speaker in a test set, or the degree-of-difficulty, varies widely, it may be more appropriate to consider properties of the ensemble of speakers. NIST's scoring software, and our tabulations of results, provide both the mean word error rate (and its associated standard deviation) and the median error rate for that ensemble of speakers.

Table 1 shows some detailed results for one system, for the entire Hub-4 evaluation test set. Each row of the table presents all of the data for each speaker in the entire test set. For some speakers, the data was obtained throughout the course of several broadcasts. Others appeared in only one broadcast. In this table, the first line presents data for an unknown speaker, the second line for "brancaccio" (David Brancaccio, the Marketplace "anchor"), the third line a speaker named "nelson", etc. The data appearing in each cell of this table consists of the aggregate number of words spoken (word tokens) by that speaker in all segments in all broadcasts, as appropriate, in parentheses, and the word error rate for that speaker's subset of the test material.

Toward the bottom of the table, the row labeled "Set Sum/Werr" presents summary numbers of word tokens, in parentheses, and the test set word error rate (expressed as a %), defined as the total number of word errors (substitutions, deletions, and insertions) divided by the number of word occurrences in that test set's (or subset's) material. Note that for the "Overall" column, the table indicates that there were 12,558 word tokens in the evaluation test set, and the corresponding test set word error rate was 42.7%.

Because error rates for heterogeneous test sets can be misleading, the bottom three rows in Table 1 present data relevant to the ensemble of speakers in the test material. The rows labeled "mean" and "StDev" present, for each set or subset, the mean (of the

number of reference transcription words spoken/speaker, in parentheses, and the mean (of the) word error rate/speaker (along with, on the lower row, the associated standard deviations) where the mean is taken over the set of all speakers.

Note, for example, that although the test set word error rate was 42.7%, the mean word error rate was considerably higher (70.5%) with a large standard deviation (51.8%). This is because there were a large number of speakers for which high error rates were found. The amount of spoken material from each speaker varies appreciably -- while the mean number of words spoken by each speaker was 348 words, the associated standard deviation was 518 words. The final row presents the medians for these data. The median word error rate is 53.7%. The mean and median error rates typically differ significantly from the test set word error rates because the distribution of the number of words, and degree of difficulty presented to the ASR technology, in the test material is non-uniform, with much of the material having been spoken by the anchor and regular correspondents.

Several partitionings of the results have been suggested and made possible using annotations of the test data.

The first shown in this table is that for sex: there are columns for the results for male and female speakers. For example, referring to the Set Sum/Avg row, it is shown that there were 1437 word tokens spoken by female speakers, in contrast to 11,121 word tokens spoken by male speakers. The corresponding word error rates are, in this case, 41.6% and 42.9%.

Another partitioning relates to "Speaking Style", and it includes columns for the "Anchor or Correspondent", "Other Am. English", and "Foreign Accented English" subsets. These categorizations are to some degree arbitrary. The first includes what may well be "read" speech by studio-based professional announcers, and regular correspondents. The "Other Am. English" category includes speech that is less formal, has more evidence of spontaneity (as is frequently found in interviews), and appears not to be "read" speech. In some cases, of course, the anchor and correspondents engage in what appears to be spontaneous dialogues, including some banter. It is difficult to categorize this. Again referring to the Set Sum/Avg row, it is shown that the "Anchor or Correspondent" subset contains 7215 word tokens, and the corresponding subset word error rate was 33.2%. In contrast, the "Other Am. English" subset had fewer word tokens (3719), and a notably higher error rate (59.2%), while the "Foreign Accent English" subset had even fewer word tokens (1624) and an error rate of 47.2%.

Another partitioning made note of the presence of "background music". For this partitioning, two columns show the data for "speech only", or "speech + music" (speech with the presence of background music in at least a portion of the relevant segment). Again referring to the same row as previously, note that, in this case, there were 11,002 word tokens in the "speech only" subset, with an error rate of 43.7%, and 1556 word tokens in the material with background music, but that the error rate for this subset (36.1%) was in fact slightly less than that for speech only. In general, this is not the case -- the presence of music is generally shown to degrade performance.

Finally, yet another partitioning of the results is possible, taking note of observations about apparent channel bandwidth. Note in this case that for the 9405 word tokens in the "full bandwidth" subset, the average word error rate was 32.9%, vs. the markedly higher word error rate of 71.9% for the 3153 word tokens in the "reduced bandwidth" subset.

Table II presents results for all officially scored results for eight "systems" for which results were submitted by four sites. In general, five of these systems yield comparable error rates. However, note that for three of the systems (variants of one system at one site) the error rates are notably lower. For this site's three systems, the test set word error rates were in the range 27.0% to 29.5%, the mean word error rates -- over the set of all speakers in the test set -- were approximately 50%, and the median word error rates ranged from 32.2% to 36.5%. Other properties are tabulated, as well. The reader is referred to other papers in this Proceedings for additional discussion of the properties of individual systems.

5. HUB-4 DISCUSSION

5.1 Properties of the Marketplace Radio Program

The Marketplace radio program is a daily weekday half-hour magazine-style program devoted to presenting news about business, the economy, and finance in an approachable manner. The program addresses a wide variety of topics (some of which are only indirectly business-related.) It employs commentary, interviews, and a variety of reporting modes. The program has an anchor, regular "staff" reporters, correspondents in several "bureaus" around the world, freelance reporters, and special commentators. There are also a few reporters who appear on a weekly or bi-weekly basis. The program is broken up into several individual stories. Unless the stories are reported by the anchor, they are each usually covered by different reporters. Feature stories usually include interviews with experts, "men on the street", sound bytes, and sound effects. The program also includes periodic special features such as "Customers from Hell", "Nashville Waitress", "Savvy Traveler". Music is used at the beginning and end of the show, during promotional segments, and at story breaks. The music is faded in and out and is quite prominent between stories.

The reporting style ranges from formal "desk" reporting to chatty interviews. The program appears to use the following format:

1. Introduction with "sound bite" from a story later in program and sponsor acknowledgments
2. Three to six feature stories
3. "Let's do the numbers", a stock market summary with either "Stormy weather" (bear market) or "We're in the money" (bull market) playing in the background
4. Sponsor acknowledgments and promotions
5. One to three additional feature stories
6. Commentary and/or a special feature (Nashville waitress, Savvy traveler, etc.)
7. News summary
8. Credits

5.2 Signal Processing for Broadcast and SNR

Traditionally, research in large-vocabulary continuous speech recognition (LV CSR) has made use of digital recordings of speech materials collected in a relatively quiet stationary noise environment, using close-talking head-mounted noise-canceling microphones, often with speaker and/or session changes known. No signal processing is implemented during data collection, other than, in some cases, "gain" adjustments to account for varying vocal effort. These recordings typically display energy histograms with discernable low-amplitude peaks, and signal-to-noise ratios in the vicinity of 40 dB are common. Figure 1(a) shows an energy histogram that is representative of traditional LV CSR corpora. Note the peak in this distribution corresponding to the background noise, approximately 40 dB below the peak speech level.

In contrast, in the course of producing radio and television broadcasts, a great deal of potentially sophisticated (digital) signal processing devices and processes are likely to have been used, including equalizers, reverberation and delay units, compressors and limiters, noise gates and expanders, etc.. These are used by the producers and broadcast engineering staffs, in some cases, to maximize the impact and commercial market for their programs, and in other cases to compensate for undesirable properties of the source materials (e.g., to improve speech intelligibility through equalization or the use of noise gates), or to enhance the vocal appeal of an announcer. The use of peak limiters and compressors, and when appropriate, noise gates, is widespread (although used to lesser degrees with broadcasts of "classical" music), with the consequence of yielding energy histograms that differ markedly from traditional LVCSR corpora, and which vary appreciably throughout the course of a broadcast. The engineering staffs of commercial stations are often reluctant to reveal the properties of their preferred signal processing procedures, since there may be potential commercial value to these "tricks of the trade".

Two additional relatively short-term energy histograms are included in Figures 1(b) and 1(c). The data in Figure 1(b) was derived from a segment of apparently read speech spoken in a studio by the "anchor", David Brancaccio. Note that the contribution to the energy histogram due to noise is minimal, and ~35 dB down from the level corresponding to the speech peak power. Figure 1(c), in contrast, was derived from a segment with British-accented, apparently read, speech over competing speech in the background. In figure 1(c), although the NIST SNR software identified a noise level ~43 dB down from the speech level, a broad peak probably corresponding to the competing speech can also be noted about 20 dB down.

Figure 1(d) illustrates a long-term energy histogram for the entire broadcast of 950801. In Figure 1(d), note that the dynamic range is, in general compressed, and note also the steep negative slope of the histogram at high levels, due to the functioning of peak limiter(s). In the long term, typically, there is no clearly identifiable "noise" peak, since broadcasters take great pains to avoid broadcasting "noise".

In general, it is meaningless to speak of "the signal-to-noise-ratio" for this test material because it varies so widely. Indeed, the variability in apparent signal-to-noise ratio can provide valuable information for segmentation of the test material.

Because of this use of signal processing, it may be the case that "off-the-air" recordings have properties that differ from those of the "master" recordings, with regard to automatic speech recognition. NIST staff arranged for a local audio contractor to perform "moderate" compression and peak limiting operations on one "Marketplace" broadcast provided on DAT from KUSC. The net effect of these operations was to compress the already limited dynamic range slightly. A limited-scale experiment, using NIST's version of Sphinx-II with this material did not seem to indicate any significant effect on word error rates. Thus it would appear that the use of materials derived from studio master tapes is equivalent to the use of "off-the-air", high quality channel broadcasts, for the purposes of this research, certainly considering the present state-of-the-art.

5.3 Phenomena Observed in the Results

A convenient way to view the results for these studies is in the form of a "bar graph" showing segment word error rate vs. time throughout the course of a broadcast (or segment), ideally with a different color used for each speaker, and with the bar widths proportional to the segment duration. Figures 2 through 6 show the segment error rates obtained with one of the well-performing systems for the entire broadcast of 950801 (Figure 2), the "head" of the 950811 broadcast (Figure 3), the 950814 "tail" (Figure 4), the 950824 "tail" (Figure 5), and the 950830 "head" (Figure 6). High error rates can be noted in the head of the 950811 broadcast -- these higher error rates for this system (and other systems) are probably attributable to a relatively large proportion of foreign-accented speech and reduced-bandwidth speech.

"Story boundaries" are shown in these figures as vertical lines. Note, for example, that for the broadcast of 8/01/95, there are 9 identified "stories" throughout the ~1700 second broadcast, the first of which starts at approximately 85 seconds into the broadcast.

The "anchor" person, (David) Brancaccio (shown in blue), appears at many times throughout the broadcast, with word error rates ranging in individual segments from a low of ~5% at ~240 seconds into the broadcast, to a high of ~40% at ~1210 seconds. The first segment introduced a story about deregulation of telecommunications, and the segment at ~1210 seconds involves apparently spontaneous speech in an interview with (Bryce) Nelson.

In this broadcast, the lowest error rate, of ~3%, was noted with the speaker named (John) Dimsdale at ~270 seconds into the broadcast, just prior to the segment involving (President) Clinton, with a word error rate of ~26.3% for the segment including the President.

Note also that the durations of the segments vary considerably, as indicated in bars of differing widths. The material involving President Clinton, for example, lasted 12.6 seconds, and included only 38 reference words.

Since many of the "stories" involve interviews, and many of the interviews are with (remote) correspondents, with the correspondent's (or interviewee's) speech transmitted to the broadcast producers' studios over reduced-bandwidth lines, error

rates for the correspondents or interviewees are typically higher than for the material originating in the studio. Note, for example, the differences in error rates for stories 3 and 7 in the broadcast of 950801, involving Brancaccio in both stories, and Barber in story 3 and Nelson in story 7.

Factors contributing to these variabilities include, of course, differences in speaking style ranging from clearly "read" speech for headline news, to banter in some of the interviews. These are in addition to differing degrees to which the material might be regarded as "business news".

5.4 System Properties

Four different sites participated in these tests -- BBN, CMU, Dragon Systems, and IBM.

The systems developed at these sites bear many common attributes, including, in most cases, the development and use of classifiers from the (limited amount of) training data, automatic segmentation ("chopping") and classification of the test segments into "acoustical categories", and channel and gender compensation, and the use of multiple passes.

Training of the language models was from a number of different sources, including: Wall Street Journal 1992-1994, North American News 1994-1995, transcriptions of Broadcast News 1992-1995 (commercially available, not including "Marketplace") broadcasts, and the transcriptions of the 10-broadcast Marketplace training material. The individual amounts of material in each of these resources ranged from ~50K words (for the 10-broadcast Marketplace training material) to ~118 M words for the commercially available Broadcast News material. Both static and adaptive language models were used, typically with approximately 50K words.

Classification schemes included detection of noisy or reduced bandwidth speech. Speaker-specific acoustic models, and the use of speaker-identification procedures made it possible to adapt acoustic models for frequently-appearing speakers (i.e., the anchor person, David Brancaccio and regular correspondents).

Because these "dry run" tests involved differing levels-of-effort and available resources at the four different sites, it is probably unreasonable to emphasize differences in performance between systems. Nonetheless, it is noteworthy that systems at three sites were closely comparable, and one site achieved markedly lower error rates.

6. ACKNOWLEDGMENTS

The authors greatly appreciate assistance from the staff of the LDC, especially Jack Godfrey and Rebecca Finch for obtaining rights to use, and obtaining "Marketplace" data, and for arranging to procure initial transcriptions from Cambridge Transcribers. At NIST, John Garofolo and Bill Fisher designed the broadcast news transcription convention. Bill Rose, Mark Przybocki, Sean Sell and others worked with John Garofolo and Jon Fiscus to verify and correct transcriptions and add additional information. Sean and Mark were responsible for down sampling the DAT tapes provided

by KUSC and making cassette copies for the transcribers. We also acknowledge with gratitude the assistance provided by several individuals at participating sites who provided preliminary sets of results for use in developing and debugging the scoring software used for these tests. Finally, P. S. Gopalakrishnan at IBM is responsible for the suggestion that tables of the form used for Table I might be informative.

NOTICE

THIS PAPER IS A DRAFT VERSION AND AS SUCH IS SUBJECT TO CHANGE PRIOR TO THE FINAL PUBLICATION IN THE PROCEEDINGS OF THIS WORKSHOP.

The results cited in Tables 1 and 2 are provided in order to stimulate analysis and discussion within the research community. Note that individual sites and systems are not identified in this table. In a previous ARPA-sponsored "dry run" test, it was noted that "certain limitations are to be imposed on the dissemination of [the] results"... including a restriction that "each participating site is permitted to use their own results for any purpose, but they cannot make use of (e.g., publish or disseminate) other sites' results without getting written permission from the sites concerned". This "dry run" was conducted as the first in what is hoped will be a series of CSR tests involving the use of radio (and possibly television) broadcasts. The data on error rates are provided as a "starting point" to document the state-of-the-art at this time -- with the use of limited and varied resources at the several participating sites.

The views expressed in this paper are those of the author(s). The results presented are for local, system-developer-implemented tests. NIST's role in these tests was one of selecting, processing and distributing the "Marketplace" broadcast materials used for training, development test and evaluation test materials, developing and implementing scoring software, and uniformly tabulating the results. The views of the author(s), and these results, are not to be construed or represented as endorsements of any systems or official findings on the part of NIST, ARPA, or the U.S. Government.

REFERENCES

- [1] Rudnicky, A.I., (Hub-4 overview paper in this proceedings)
- [2] (Hub-4 BBN paper in this proceedings)
- [3] (Hub-4 CMU paper in this proceedings)
- [4] (Hub-4 Dragon Systems paper in this proceedings)
- [5] (Hub-4 IBM paper in this proceedings)
- [6] See, for example, Section 14 "Signal Processing Equipment" in "The Sound Reinforcement Handbook", second edition, by Gary Davis and Ralph Jones, published by Hal Leonard Corp., Milwaukee, WI, 1989. ISBN 0-88188-900-8.

APPENDIX 1.

Example portion of transcription for Marketplace 08/01/95 broadcast:

<broadcast id="marketplace.950801" rev="951116">

[music/] [time=0.90]

A(bt=5.26 et=7.97): From Los Angeles, this is Marketplace.

B(bt=19.05 et=29.10): Trying not to be outdone by the Disney A\ B\ C\ buy-out, C\ B\ S\ and Westinghouse announce a deal of their own. Will major media mergers make life harder for journalists?

C(bt=29.16 et=44.23): The whole kind of consolidation of media ownership means that uh generally reporters probably become less courageous in reporting o- on business uh because they uh who knows when your next uh who your next owner is going to be?

B(bt=44.60 et=50.41): And the savvy traveler's found a couple of exotic getaways that are almost affordable. This is Marketplace.

A(bt=54.72 et=76.37): Marketplace is produced by K U S C at the University of Southern California for Public Radio International and is made possible by G\ E\ . From aircraft engines to appliances to broadcasting, G\ E\ , we bring good things to life. And by the Corporation for Public Broadcasting and public radio stations nationwide.

{{interlude}}

B(bt=79.49 et=106.73): It's Tuesday, August first. I'm David Brancaccio and here's some of what's happening in business and the world.

<story id=1 topic="CBS buy-out" bt=84.98 et=233.46>

[inhaling] Call it the Consolidated Broadcasting System. Upstaged by yesterday's huge Disney and Capl. Cities agglomeration, the long-awaited Westinghouse bid for the C\ B\ S\ network finally came down late today [inhaling] in a deal worth almost five and a half billion dollars. [inhaling] Marketplace's Philip Boroff is at the Waldorf Astoria hotel in New York where the agreement was announced. [inhaling] Philip, [music/] [time=105.77] what are the terms?

[audio_change]

D(bt=106.85 et=143.08): Well, Westinghouse is offering eighty one dollars per share which adds up to five point four billion dollars which, of course, sounds like a lot, [inhaling] but it's less than a third of the deal yesterday. [inhaling] And I thought the tone of the press conference today was sort of defensive. Yesterday [talking/] Michael Eisner was real confident, but here uh Michael Jordan, the chief of Westinghouse, no relation to the athlete, [inhaling] I felt was real defensive. He talked a lot about tax benefits and operating margins and cash flow [inhaling] and it sounds more like one of those traditional mergers where you put

the two companies together. Then you save money by cutting jobs. He didn't say there would definitely be job cuts but he certainly refused to uh rule them out. [/talking]

[audio_change]

B(bt=143.10 et=156.11): Philip, there were questions today about whether C\ B\ S\ chairman Lawrence @Tish might be hedging his bets a little here. He personally owns, after all, a huge chunk of C\ B\ S\ and [inhaling] some were apparently wondering if his stock is now inexorably pledged to Westinghouse.

[audio_change]

D(bt=156.24 et=167.85): [inhaling] He was asked about that today and he kind of skirted around the question. He said [background_talking/] he pledged his love to Mr\ Jordan, the Westinghouse C\ E\ O\ , but he also said he had a fiduciary obligation to consider other bids. [/background_talking]

[audio_change]

B(bt=167.96 et=171.23): Other bids, very interesting. Philip Boroff in New York, thank you very much.

[audio_change]

D(bt=171.23 et=172.10): Thank you.

[audio_change]

B(bt=172.14 et=185.99): So what does a diversified industrial concern like Westinghouse hope C\ B\ S\ can do for its bottom line? Martin @Piers, a business and finance writer at Variety, says Westinghouse C\ E\ O\ Michael Jordan has been charged with revitalizing the company.

[audio_change]

E(bt=186.18 et=231.75): As I understand it, he realized fairly quickly that one of Westinghouse's best performing businesses was its broadcasting division. [inhaling] Um you know, they've been in radio since the hist(ory)- since radio began. [inhaling] Um they're also uh quite big in the television station business and they've got a programming business which hasn't done all that well but they've been in that for quite a quite a l- long time as well. So anyway, he realized that that performed very well. It's um I think it contributes about a third of the cash flow of the overall company [inhaling] and ((I)) (()) you know it seems that he decided that um maybe the thing to do is to try (()) t- t- to turn around Westinghouse was to concentrate more on broadcasting and less on some of these other troubled industrial businesses.

[audio_change]

B(bt=231.84 et=233.41): Martin @Piers at Variety.

</story>

APPENDIX 2.

Example portion of speaker information for Marketplace 08/01/95 broadcast

```
{{speaker info for marketplace.950801 rev="951116"}}
```

```
speaker_a_name: unknown
speaker_a_role: announcer
speaker_a_sex: female
```

```
speaker_b_name: David Brancaccio
speaker_b_role: Marketplace anchor
speaker_b_sex: male
```

```
speaker_c_name: Bryce Nelson
speaker_c_role: former Chicago bureau chief, LA Times
speaker_c_sex: male
{this speaker is nearly incomprehensible...}
```

```
speaker_d_name: Philip Boroff
speaker_d_role: Marketplace Reporter
speaker_d_sex: male
```

```
speaker_e_name: Michael @Piers
speaker_e_role: Variety writer
speaker_e_sex: male
speaker_e_dialect: British
```

```
speaker_f_name: John Dimsdale
speaker_f_role: Marketplace Washington editor
speaker_f_sex: male
```

```
speaker_g_name: Bill Clinton
speaker_g_role: US President
speaker_g_sex: male
```

```
speaker_h_name: Benjamin Barber
speaker_h_role: Rutgers Political Science Professor
speaker_h_sex: male
```

```
speaker_i_name: unknown
speaker_i_role: singers
speaker_i_sex: female and male
```

```
.
.
.
```

APPENDIX 3.

In most automatic speech recognition benchmark tests implemented by NIST, the primary performance measure is of word error rate, expressed as a percentage of the number of word tokens in the reference material. Assuming that some string-matching process has been used to align both reference transcriptions and system "hypothesis" transcriptions, errors of three types are identified and counted: substitutions (#subs), deletions (#dels), and insertions (#ins). The (unweighted) sum of these is regarded as the total number of word errors.

Then, the word error rate expressed as a percentage is

$$werr = (100) \frac{\#subs + \#dels + \#ins}{\#refwords}$$

Given a test set (or subset) t comprising S speakers each speaking a number of "utterances" (or, in some cases, as in the Hub-4 tests, "segments of speech"), the above expression for the test set word error rate may be written

$$werr(t) = (100) \frac{\sum_{s \in t} \sum_{u \in S} (\#subs(u) + \#dels(u) + \#ins(u))}{\sum_{s \in t} \sum_{u \in S} \#refwords(u)}$$

In some cases, it is of interest to determine the word error rate for an individual speaker, $werr(s)$. In this case

$$werr(s) = 100 \frac{\sum_{u \in S} (\#subs(u) + \#dels(u) + \#ins(u))}{\sum_{u \in S} \#refwords(u)}$$

The mean speaker word error rate over the set of speakers s in a test set t is

$$mean\ speaker\ werr(t) = \overline{werr(s)} = \frac{\sum_{s \in t} werr(s)}{S}$$

and the associated sample standard deviations is

$$Std.\ Dev.\ werr(t) = \frac{\sum_{s \in t} (werr(s) - \overline{werr(s)})^2}{S-1}$$

In some cases, the median word error rate over the set of speakers is of interest:

$$median\ speaker\ werr(t) = median\{werr(s) \mid s \in t\}$$

Overall
 Female
 Male
 Anchor or Correspondent
 Other Am English
 Foreign Accents
 Speech only
 Speech+Music
 Full BW
 Reduced BW

Overall
 Female
 Male
 Anchor or one of the correspondents
 Other Spanish or English
 Foreign Accents
 Segments containing only speech
 Segments containing speech and music
 Segments which contain full bandwidth speech
 Segments with reduced bandwidth speech

SPER	Overall		Female		Male		Anchor or Correspondent		Speech Style		Background Music		Channel Bandwidth					
	BW	WZ	BW	WZ	BW	WZ	BW	WZ	Other Am English	Foreign Accents	Speech only	Speech+Music	Full BW	Reduced BW				
unknown017	[318]	30.6	[216]	30.6	[218]	30.6	[218]	30.6	[189]	53.1	[190]	52.9	[191]	56.4	[199]	53.1	[172]	74.4
braccio	[3064]	26.0	[204]	26.0	[204]	26.0	[204]	26.0	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4
nelson	[589]	53.1	[589]	53.1	[589]	53.1	[589]	53.1	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6
boyer	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4	[172]	74.4
piers	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6	[134]	80.6
dimitrie	[178]	74.4	[178]	74.4	[178]	74.4	[178]	74.4	[138]	50.0	[138]	50.0	[138]	50.0	[138]	50.0	[138]	50.0
clinton	[38]	50.0	[38]	50.0	[38]	50.0	[38]	50.0	[480]	84.2	[480]	84.2	[480]	84.2	[480]	84.2	[480]	84.2
barber	[480]	84.2	[480]	84.2	[480]	84.2	[480]	84.2	[5]	120.0	[5]	120.0	[5]	120.0	[5]	120.0	[5]	120.0
unknown018	[5]	120.0	[5]	120.0	[5]	120.0	[5]	120.0	[304]	18.4	[304]	18.4	[304]	18.4	[304]	18.4	[304]	18.4
lennedy	[424]	22.6	[424]	22.6	[424]	22.6	[424]	22.6	[280]	92.5	[280]	92.5	[280]	92.5	[280]	92.5	[280]	92.5
June	[280]	92.5	[280]	92.5	[280]	92.5	[280]	92.5	[308]	39.0	[308]	39.0	[308]	39.0	[308]	39.0	[308]	39.0
dube	[308]	39.0	[308]	39.0	[308]	39.0	[308]	39.0	[322]	32.2	[322]	32.2	[322]	32.2	[322]	32.2	[322]	32.2
stichole	[292]	32.2	[292]	32.2	[292]	32.2	[292]	32.2	[499]	49.7	[499]	49.7	[499]	49.7	[499]	49.7	[499]	49.7
mare	[499]	49.7	[499]	49.7	[499]	49.7	[499]	49.7	[6]	100.0	[6]	100.0	[6]	100.0	[6]	100.0	[6]	100.0
unknown028	[6]	100.0	[6]	100.0	[6]	100.0	[6]	100.0	[134]	79.1	[134]	79.1	[134]	79.1	[134]	79.1	[134]	79.1
meyer	[134]	79.1	[134]	79.1	[134]	79.1	[134]	79.1	[489]	52.0	[489]	52.0	[489]	52.0	[489]	52.0	[489]	52.0
johnson	[489]	52.0	[489]	52.0	[489]	52.0	[489]	52.0	[16]	268.8	[16]	268.8	[16]	268.8	[16]	268.8	[16]	268.8
unknown031	[16]	268.8	[16]	268.8	[16]	268.8	[16]	268.8	[210]	53.1	[210]	53.1	[210]	53.1	[210]	53.1	[210]	53.1
beard	[210]	53.1	[210]	53.1	[210]	53.1	[210]	53.1	[64]	82.8	[64]	82.8	[64]	82.8	[64]	82.8	[64]	82.8
mliter	[64]	82.8	[64]	82.8	[64]	82.8	[64]	82.8	[116]	97.4	[116]	97.4	[116]	97.4	[116]	97.4	[116]	97.4
gordon	[116]	97.4	[116]	97.4	[116]	97.4	[116]	97.4	[33]	81.8	[33]	81.8	[33]	81.8	[33]	81.8	[33]	81.8
titelli	[33]	81.8	[33]	81.8	[33]	81.8	[33]	81.8	[9]	222.2	[9]	222.2	[9]	222.2	[9]	222.2	[9]	222.2
unknown030	[9]	222.2	[9]	222.2	[9]	222.2	[9]	222.2	[198]	54.0	[198]	54.0	[198]	54.0	[198]	54.0	[198]	54.0
giles	[198]	54.0	[198]	54.0	[198]	54.0	[198]	54.0	[820]	29.8	[820]	29.8	[820]	29.8	[820]	29.8	[820]	29.8
curtman	[820]	29.8	[820]	29.8	[820]	29.8	[820]	29.8	[382]	42.9	[382]	42.9	[382]	42.9	[382]	42.9	[382]	42.9
garzanti	[382]	42.9	[382]	42.9	[382]	42.9	[382]	42.9	[895]	33.7	[895]	33.7	[895]	33.7	[895]	33.7	[895]	33.7
kaussak	[895]	33.7	[895]	33.7	[895]	33.7	[895]	33.7	[458]	39.1	[458]	39.1	[458]	39.1	[458]	39.1	[458]	39.1
shorer	[458]	39.1	[458]	39.1	[458]	39.1	[458]	39.1	[6]	100.0	[6]	100.0	[6]	100.0	[6]	100.0	[6]	100.0
garcher	[609]	20.7	[609]	20.7	[609]	20.7	[609]	20.7	[22]	86.4	[22]	86.4	[22]	86.4	[22]	86.4	[22]	86.4
goldman	[329]	51.4	[329]	51.4	[329]	51.4	[329]	51.4	[217]	93.5	[217]	93.5	[217]	93.5	[217]	93.5	[217]	93.5
sealy	[16]	100.0	[16]	100.0	[16]	100.0	[16]	100.0	[731]	98.4	[731]	98.4	[731]	98.4	[731]	98.4	[731]	98.4
seaweb	[22]	86.4	[22]	86.4	[22]	86.4	[22]	86.4	[353]	16.1	[353]	16.1	[353]	16.1	[353]	16.1	[353]	16.1
goldberg	[217]	93.5	[217]	93.5	[217]	93.5	[217]	93.5	[27]	100.0	[27]	100.0	[27]	100.0	[27]	100.0	[27]	100.0
farrell	[731]	98.4	[731]	98.4	[731]	98.4	[731]	98.4	[4121]	41.4	[4121]	41.4	[4121]	41.4	[4121]	41.4	[4121]	41.4
neeson	[353]	16.1	[353]	16.1	[353]	16.1	[353]	16.1	[303]	70.3	[303]	70.3	[303]	70.3	[303]	70.3	[303]	70.3
unknown037	[27]	100.0	[27]	100.0	[27]	100.0	[27]	100.0	[518]	51.8	[518]	51.8	[518]	51.8	[518]	51.8	[518]	51.8
Set Run/Waz	[12598]	42.7	[12598]	42.7	[12598]	42.7	[12598]	42.7	[216]	91.7	[216]	91.7	[216]	91.7	[216]	91.7	[216]	91.7
Mean	[303]	70.3	[303]	70.3	[303]	70.3	[303]	70.3	[197]	79.1	[197]	79.1	[197]	79.1	[197]	79.1	[197]	79.1
AtDw	[518]	51.8	[518]	51.8	[518]	51.8	[518]	51.8	[418]	45.6	[418]	45.6	[418]	45.6	[418]	45.6	[418]	45.6
Medica	[216]	91.7	[216]	91.7	[216]	91.7	[216]	91.7	[197]	79.1	[197]	79.1	[197]	79.1	[197]	79.1	[197]	79.1

Table 1.

For the Nov85 ARPA CSR HA "Dry Run" Evaluation Test

Overall
 Female Speaker
 Male Speaker
 Anchor or correspondent
 Other speakers of U.S. English
 Other speakers of U.S. English
 Speakers with foreign accent
 Segments containing only speech
 Segments containing speech and music
 Segments which contain full bandwidth speech
 Segments with reduced bandwidth speech

Overall
 Female
 Male
 Anchor or correspondent
 Other speakers of U.S. English
 Foreign Accent Eng.
 Speech only
 Speech+Music
 Full BW
 Reduced BW

SYSTEM	Overall		Female		Male		Speaking style		Background Music		Channel Bandwidth	
	Words	Rate	Words	Rate	Words	Rate	Speech only	Speech+Music	Full BW	Reduced BW	Words	Rate
012581	45.7	(1437)	41.6	(1121)	42.9	(1215)	28.0	(3719)	57.6	(1624)	54.6	(1502)
012582	41.1	(1437)	31.2	(1121)	42.4	(1215)	30.0	(3719)	57.0	(1624)	54.5	(1502)
012583	41.4	(1437)	30.3	(1121)	42.8	(1215)	29.7	(3719)	60.6	(1624)	49.4	(1502)
012584	41.6	(1437)	30.3	(1121)	43.3	(1215)	30.0	(3719)	60.3	(1624)	49.2	(1502)
012585	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012586	39.5	(1437)	30.5	(1121)	40.8	(1215)	20.8	(3719)	31.5	(1002)	29.2	(1002)
012587	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012588	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012589	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012590	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012591	37.0	(1437)	27.9	(1121)	37.9	(1215)	18.8	(3719)	37.5	(1002)	27.0	(1002)
012592	70.5	(205)	66.3	(383)	66.2	(555)	39.8	(322)	98.6	(232)	63.1	(333)
012593	65.7	(205)	75.4	(383)	63.4	(555)	27.9	(322)	96.4	(232)	66.9	(333)
012594	64.8	(205)	78.2	(383)	61.6	(555)	28.5	(322)	92.6	(232)	63.6	(333)
012595	63.6	(205)	71.9	(383)	61.6	(555)	29.2	(322)	91.6	(232)	63.6	(333)
012596	49.9	(205)	68.7	(383)	45.4	(555)	18.4	(322)	74.9	(232)	31.4	(333)
012597	51.5	(205)	69.4	(383)	47.3	(555)	19.5	(322)	76.8	(232)	33.0	(333)
012598	49.9	(205)	68.6	(383)	45.4	(555)	18.2	(322)	74.9	(232)	31.6	(333)
012599	49.9	(205)	68.6	(383)	45.4	(555)	18.2	(322)	74.9	(232)	31.6	(333)
012600	49.9	(205)	68.6	(383)	45.4	(555)	18.2	(322)	74.9	(232)	31.6	(333)
012601	51.8	(218)	70.9	(565)	46.6	(763)	22.7	(267)	62.6	(224)	24.3	(438)
012602	57.5	(218)	71.9	(565)	44.7	(763)	11.9	(267)	72.4	(224)	37.2	(438)
012603	54.5	(218)	79.4	(565)	48.0	(763)	13.3	(267)	68.9	(224)	35.2	(438)
012604	49.7	(218)	68.2	(565)	45.2	(763)	13.3	(267)	65.9	(224)	32.8	(438)
012605	49.0	(218)	68.2	(565)	45.2	(763)	13.3	(267)	65.9	(224)	32.8	(438)
012606	58.2	(218)	76.9	(565)	53.4	(763)	6.8	(267)	77.7	(224)	30.1	(438)
012607	58.1	(218)	76.9	(565)	53.4	(763)	7.7	(267)	78.2	(224)	30.8	(438)
012608	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012609	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012610	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012611	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012612	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012613	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012614	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012615	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012616	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012617	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012618	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012619	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012620	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012621	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012622	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012623	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012624	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012625	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012626	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012627	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012628	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012629	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012630	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012631	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012632	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012633	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012634	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012635	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012636	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012637	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012638	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012639	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012640	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012641	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012642	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012643	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012644	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012645	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012646	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012647	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012648	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012649	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012650	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012651	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012652	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012653	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012654	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012655	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012656	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012657	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012658	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012659	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012660	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012661	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012662	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012663	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012664	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012665	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012666	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012667	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012668	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012669	57.6	(218)	77.0	(565)	52.6	(763)	7.2	(267)	77.7	(224)	37.3	(438)
012670	57.6	(218)	77.0									

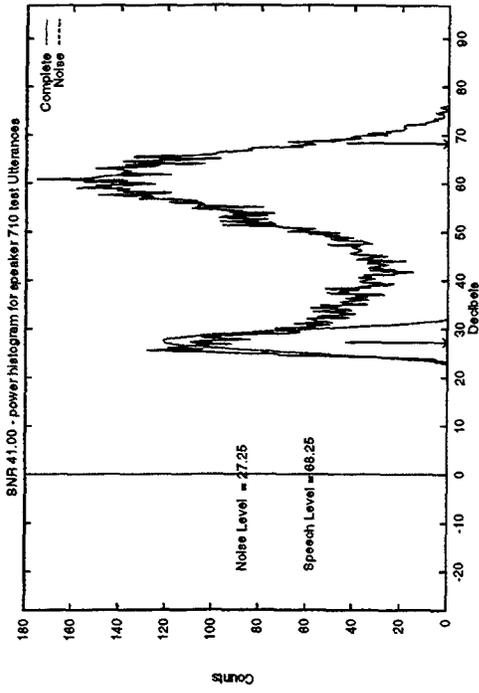


Figure 1(a).

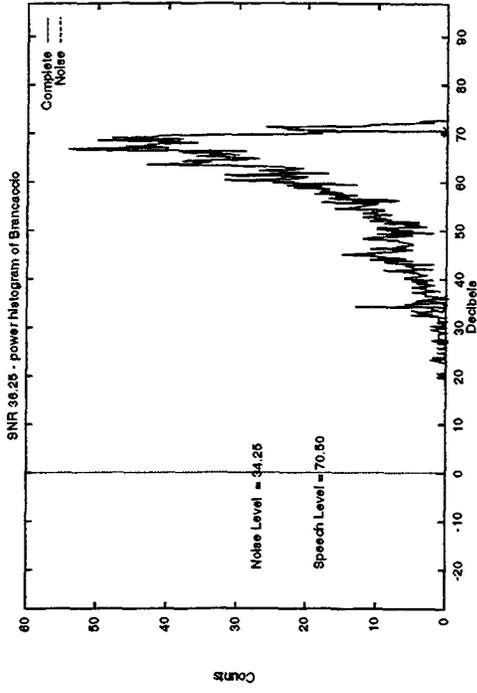


Figure 1(b).

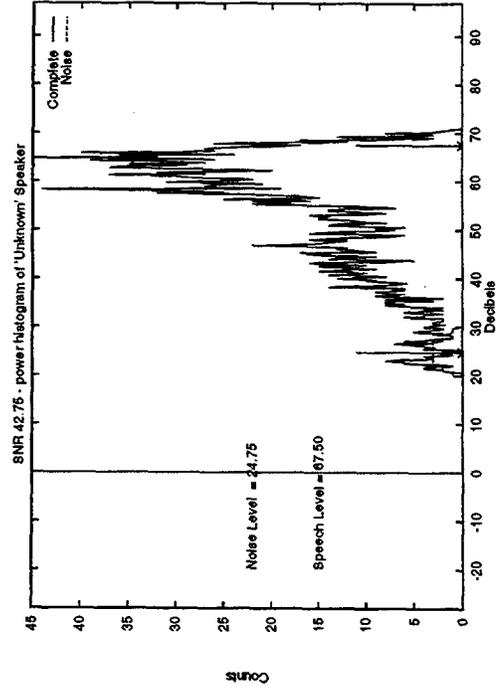


Figure 1(c).

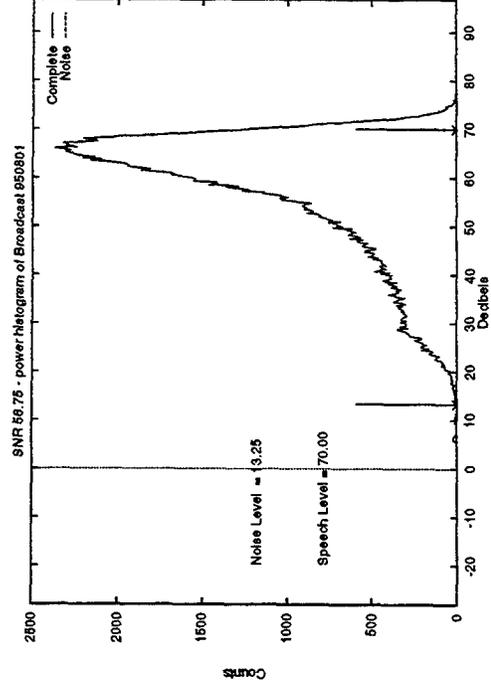


Figure 1(d).

SEGMENT ERROR RATES VERSUS TIME
 (950801 ENTIRE BROADCAST)

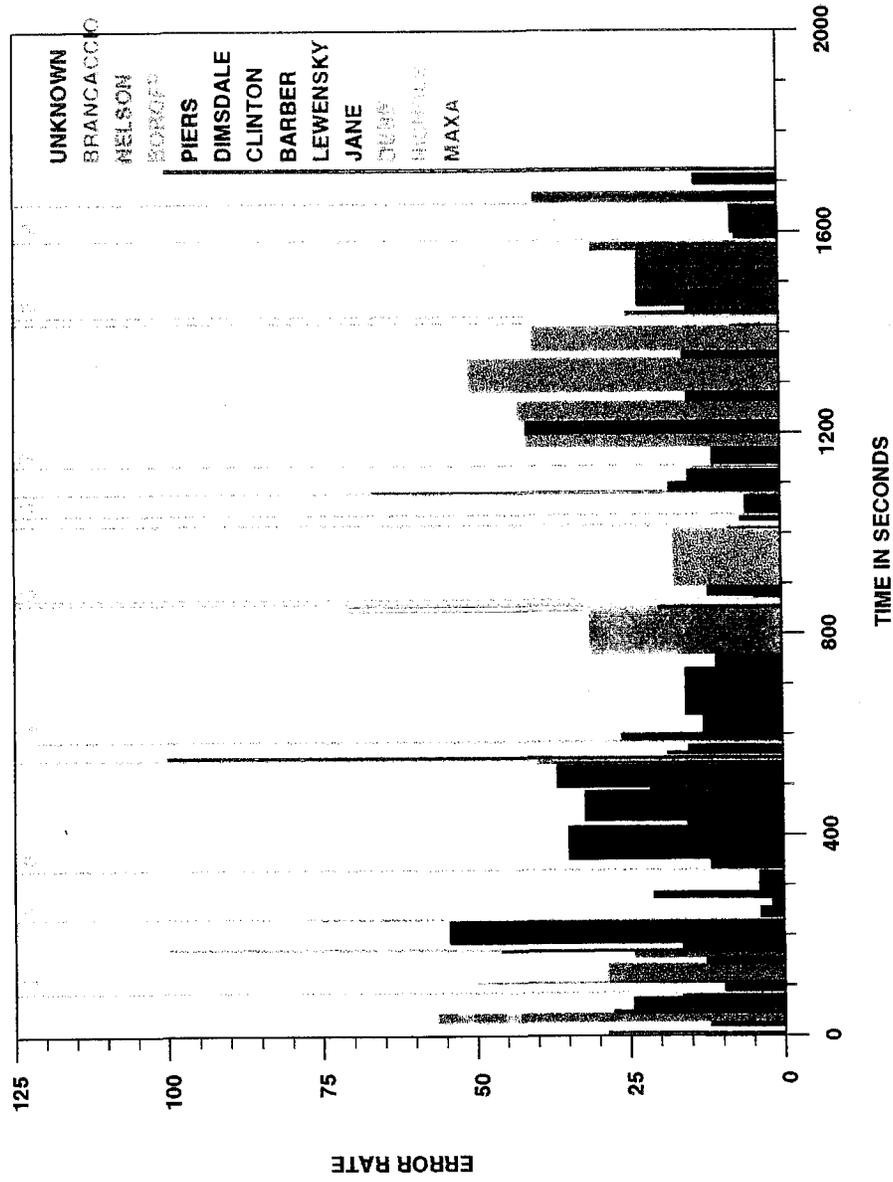


Figure 2.

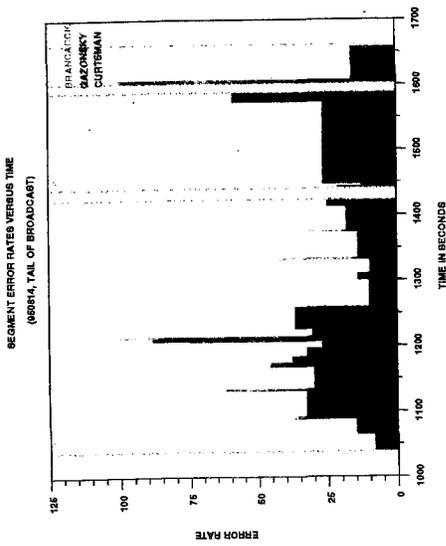


Figure 3.

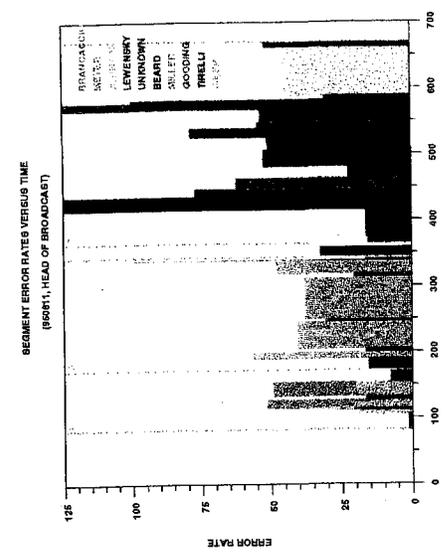


Figure 4.

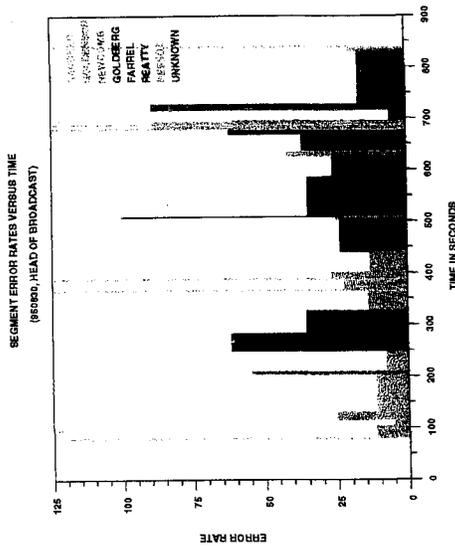


Figure 5.

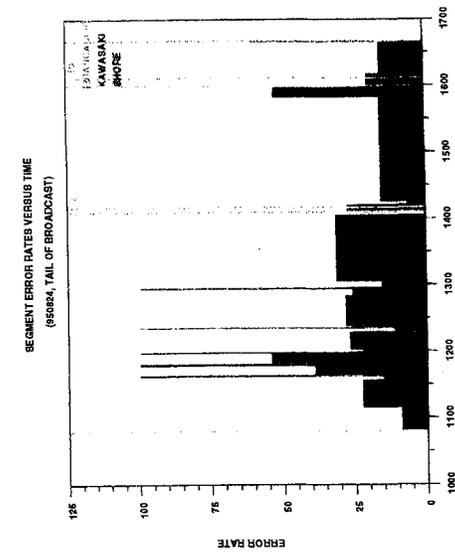


Figure 6.