

SELECTED TEST MATERIAL
FOR THE
MARCH 1987 DARPA BENCHMARK TESTS

David S. Pallett

Institute for Computer Sciences and Technology
National Bureau of Standards
Gaithersburg, MD 20899

ABSTRACT

This paper describes considerations in selecting test material for the March '87 DARPA Benchmark Tests. Using a subset of material available from the Task Domain (Resource Management) Development Test Set, two sets of 100 sentence utterances were identified. For Speaker Independent technology, 10 speakers each provide 10 test sentences. For Speaker Dependent technology, 4 speakers each provide 25 test sentences. For "live talker" test purposes, three 30-sentence scripts were identified, using a total of 70 unique sentence texts. The texts of all of these test sentences were drawn from a set of 2200 sentences developed by BBN in modelling the (resource management) task domain.

INTRODUCTION

In order to implement benchmark tests of speech recognition systems to be reported at the March '87 DARPA Speech Recognition Meeting, it was necessary to specify selected test material. This test material is drawn from two sources: (a) the Task Domain Speech Database recorded at Texas Instruments (also referred to as the "Resource Management" Database), and (b) the use of "live talkers" in site visits. In each case, the texts of the sentences were drawn from a set of sentences developed by BBN. Selection of test material using the Resource Management Database includes two separate components, a Speaker Independent component and a Speaker Dependent component. This paper outlines the process of defining these subsets of speech material.

At the time the Resource Management Speech Database was designed, it was intended that approximately equal volumes of material would be available for system

development (research) purposes and for two rounds of benchmark tests. Consequently, approximately half of the available material is designated "development" or "training" material, and the remaining portion is designated for test purposes. The test material is designated as "Development Test" or "Evaluation Test" sets, each including 1200 test sentence utterances in each portion (Speaker Independent or Speaker Dependent).

The design and collection of this Task Domain (Resource Management) Speech Database is described elsewhere in this Proceedings in a paper by Fisher [1].

Thus, as originally intended, two sets of 1200 sentence utterances were to be available for the March '87 tests. During January 1987, discussions involving representatives of CMU, BBN, MIT, NBS and the DARPA Program Manager determined that use of this large a volume of test material was not necessary to establish performance of current technology when pragmatic considerations of processing times and expected performance levels were made. Consequently, it was agreed that subsets of 100 sentence utterances were to be defined for these tests, and that NBS would specify the appropriate subset.

To complement the use of the recorded speech database material, a test protocol for the use of "live talkers" emulating in some sense procedures to be used in future demonstrations of these systems was defined, and texts were selected for this purpose.

RESOURCE MANAGEMENT SPEECH DATABASE TEST MATERIAL

Speaker Independent Test Material

For the March '87 tests, a set of ten speakers was identified, drawn from material recorded at TI and made available to NBS in December '86 and January '87.

Each speaker provided two "dialect" and the ten "rapid adaptation" sentences in addition to a total of thirty test sentence utterances. For each speaker, a unique subset of ten sentence utterances were specified to be used for the March '87 tests, amounting to 100 sentence utterances in all (10 speakers times 10 sentence utterances per speaker).

Seven male speakers were selected and three female speakers, reflecting the male/female balance throughout the Resource Management Speech Database.

To aid in the selection of individual speakers, a set of approximately 16 speakers was identified. SRI was asked for advice on whether any of these would be regarded as anomalous on the basis of the "dialect" sentences obtained in the acoustic-phonetic database. SRI performed a clustering analysis and advised us that most of the speakers clustered in three groups of similar speakers with three other individuals categorized as exceptional in some sense (e.g. unusually slow rate of speech) [2]. The ten speakers identified for inclusion in the test subset include one of these "exceptional" speakers, the others being drawn from the three clusters to provide some degree of coverage of regional effects.

Table 1 provides detailed information on the individual speakers' regional backgrounds, race, year of birth and educational level for the ten selected speakers in the March '87 Test Subset.

Analysis, by TI, of the lexical coverage provided by this subset of the test material indicates that 348 words occur at least once in this test material, and the total number of words is 836, for a mean length of each sentence of 8.36 words.

Subject	Sex	Region	Race	Year of Birth	Education
DAB	MALE	NEW ENGLAND	WHT	'62	B.S.
GWT	MALE	NORTHERN	WHT	'21	B.S.
DLG	MALE	NORTH MIDLAND	WHT	'42	(?)
CTT	MALE	SOUTHERN	WHT	'62	B.S.
JFC	MALE	NEW YORK CITY	WHT	'59	B.S.
BTH	MALE	WESTERN	WHT	'62	B.S.
AWF	FEMALE	SOUTHERN	WHT	'58	B.S.
BCG	FEMALE	"ARMY BRAT"	(?)	'59	B.S.
SAH	FEMALE	NEW ENGLAND	WHT	'46	B.S.
JFR	MALE	WESTERN	WHT	'39	M.S.

Table 1. Speaker Independent Test Subset

Speaker Dependent Test Material

For these tests, a set of four speakers was identified, also drawn from material recorded at TI and made available to NBS during December '86 and January '87. In this case, selection of the specific individuals was strongly influenced by the availability of training material. BBN expressed concern that the entire set of 600 sentence utterances intended for system training should be available for any test speakers. At the time of selection of test material, not all of the 12 speakers for this portion of the database had completed recording their training material. With this in mind four speakers were identified.

Each speaker had previously recorded the ten "rapid adaptation" and "dialect" sentences, and the Development Test material included 100 sentence utterances for each speaker. From this, unique sets of 25 sentence utterances were identified for each of the four speakers, amounting to 100 sentence utterances in all for this portion of the test material.

Three of the speakers were male and one was female.

Table 2 provides additional data on these speakers.

Analysis, by TI, of the lexical coverage provided by this subset of the test material indicates that 832 words occur at least once, with a total number of words of 832, for a mean sentence length of 8.32 words. This is quite similar to that for the Speaker Independent material, although the details of the distributions differ slightly.

Subject	Sex	Region	Race	Year of Birth	Education
CMR:	FEMALE	NORTHERN	WHT	'51	M.S.
BEF:	MALE	NORTH MIDLAND	WHT	'52	Ph.D
JWS:	MALE	SOUTH MIDLAND	WHT	'40	B.S.
RKM:	MALE	SOUTHERN	BLK	'56	B.S.

Table 2. Speaker Dependent Test Subset

LIVE TALKER TEST MATERIAL

For the "live tests", it was necessary to select sentence texts that would be read by the test speakers. It was thought desirable to use three speakers, each speaker reading a total of 30 sentence texts in addition to the 10 "rapid adaptation" sentences. Ten of the thirty sentence texts were to be the same

for all speakers, so that of the 90 sentence utterances to be used for testing, there would be three productions of each of the ten sentences, and 60 other sentences (20 for each of three speakers). A total of 70 unique sentence texts was thus required.

The sentence texts were selected from a subset of 2200 Resource Management sentences. CMU representatives had indicated a preference for sentence texts that could be produced in less than 6 seconds. Accordingly, the essentially random process of sentence text selection was perturbed slightly to throw out longer sentences.

Lexical analysis, by TI, of the scripts developed from these sentences indicates that the three scripts are well-balanced in terms of mean sentence length and number of lexical entries. Each of the three scripts has a mean sentence length of 7.93 words (258 words/30 sentences), reflecting the intentional bias in sentence selection process toward slightly shorter sentences. The number of lexical entries in the three scripts is 153, 155 and 161.

The prompt form of each of these scripts was to be made available to the "live talkers" in site visits to be conducted in March '87. Each of the test speakers was to use the Sennheiser HMD 414-6, the same microphone used at TI for the Resource Management Speech Database, and the test environment was to be a computer lab or conference room with no competing conversation. A portion of the test material was to be provided in an interactive manner (i.e. while waiting for system processing of the data) and the remainder was to be processed off line.

GRAMMATICAL COVERAGE

At the time that BBN developed the set of approximately 2800 sentence texts modelling this task domain, no explicit or formally defined grammar was used. Rather, a set of prototypical sentences was identified to provide coverage of the task, and the subset of vocabulary occurring in these sentence "patterns" was then expanded to approximately 1000 words. There were a total of approximately 950 sentence patterns [3]. By incorporation of the expanded vocabulary, the 2800 sentences were generated by including approximately three exemplars of each pattern. From these, 600 were designated to be used for speaker-dependent training material, leaving a remaining subset of 2200 sentences. All of the test material was randomly selected from this subset of 2200 sentences.

No analysis to determine the representation of the basic sentence patterns in the test material has been conducted to date.

REFERENCES

- [1] W. A. Fisher, "A Task Domain Database", Proceedings of the March 1987 DARPA Speech Recognition Workshop.
- [2] J. Bernstein, private communication, January 1987.
- [3] P. Price *et al.*, oral presentation at the September 1986 DARPA Speech Recognition Workshop.