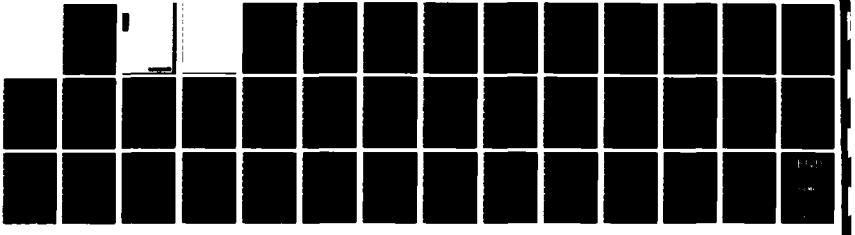
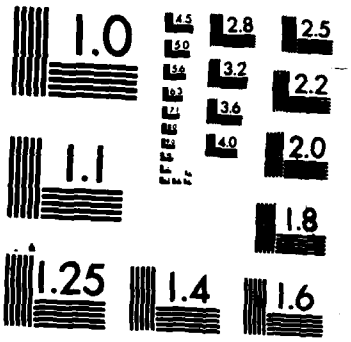
AD-A160 401 THE INTELLIGIBILITY OF NON-VOCODED AND VOCODED SEMANTICALLY ANOMALOUS SEN..(U) MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB M A MACK ET AL. 26 JUL 85 1/1

UNCLASSIFIED TR-703 ESD-TR-85-211 F19628-85-C-0002 F/G 5/7 NL





MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

THE INTELLIGIBILITY OF NON-VOCODED AND VOCODED SEMANTICALLY ANOMALOUS SENTENCES

M.A. MACK

B. GOLD

Group 24

TECHNICAL REPORT 703

26 JULY 1985

LEXINGTON MASSACHUSETTS

ABSTRACT

This document

~~The present study~~ is devoted to an analysis of the intelligibility of semantically anomalous sentences presented in four acoustically different conditions: (1) natural speech, no noise; (2) vocoded speech, no noise; (3) vocoded speech, noise added to the pitch track; (4) vocoded speech, noise added to the spectrum. One of our objectives was to analyze the specific types of errors in each condition. ~~Our~~ The other objective was to compare results of this analysis with results obtained from the Diagnostic Rhyme Test (DRT). Results revealed that intelligibility was quite good in conditions (1) and (2), relatively poor in (3), and quite poor in (4) — results consistent with DRT data. Further, about 60% of all errors were phonemic, while 40% were syntactic and semantic. We conclude that information in the spectrum is more critical than information in the pitch track, that most errors affect the phonological component when intelligibility is poor and context is uncertain, and that the DRT is an appropriate though perhaps insufficient test of speech intelligibility.

Additional keywords: Speech perception.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

THE INTELLIGIBILITY OF NON-VOCODED AND VOCODED SEMANTICALLY ANOMALOUS SENTENCES

I. INTRODUCTION

Over the past thirty to forty years, numerous experiments designed to assess speech intelligibility have been conducted. In these experiments, researchers have employed a variety of paradigms and stimuli. As a consequence, there exists a rather rich repository of information on speech intelligibility. Viewed individually and collectively, these experiments have provided valuable data regarding such factors as the perceptual effects of various signal-to-noise ratios, the contribution of contextual information to intelligibility, and the relative salience of synthetic versus natural speech.

For example, in their classic experiment, Miller and Nicely[1] found that random masking noise had more deleterious effects upon the perception of certain consonants and phonemic features than others. In a more recent experiment in word intelligibility, Kalikow, Stevens, and Elliot[2] found that the context in which words were presented bore directly upon their intelligibility. For when subjects were presented with sentences in "babble-type" noise, they consistently comprehended sentence-final words in sentences of high probability better than they did sentence-final words in sentences of low probability. And, in a phoneme-recognition experiment, subjects tested by Pisoni and Hunnicutt[3] performed better in response to natural speech than to synthetic speech.

A related issue in speech intelligibility pertains to the design of the tests themselves. In brief, there has been an ongoing attempt to identify and utilize the "ideal" testing procedure. At one end of the spectrum have been experiments in which stimuli are presented in nearly context-free situations. An example of this approach is the Diagnostic Rhyme Test or DRT. This test consists of the presentation of minimal pairs, such as "caught" and "taught," which differ only with respect to one word-initial phoneme, and it utilizes a closed forced-choice design. At the other end of the spectrum have been experiments in which stimuli are presented in relatively elaborated linguistic contexts, such as meaningful sentences.

The complexity of the test stimuli and the demands of the test are not trivial matters in the assessment of speech intelligibility. Indeed, they may affect subjects' performance significantly. A straightforward example of this has been provided in an experiment conducted by Pisoni and Koen[4] who found that subjects performed more accurately on tests of word intelligibility when they were required to respond with one of six alternatives than when they were permitted to respond freely.

Findings such as these bear directly on research designed to evaluate and improve the intelligibility of synthetic vocoded speech. For several years, the Speech Systems Technology Group at the M.I.T. Lincoln Laboratory has made use of the DRT to help assess the intelligibility of various vocoder systems. Essentially, DRT results with vocoded-speech stimuli have been consistent

with results of other experiments in which natural and/or synthetic speech has been used in comparable test paradigms. For example, Gold and Tierney[5] found that poor-quality vocoded speech, or vocoded speech presented in noise, resulted in large decrements in intelligibility, and that these decrements consistently affected certain phonemes more than others. Moreover, even extremely high-quality vocoded speech yielded somewhat lower DRT scores than did natural speech.

Further, the Gold and Tierney results revealed that greater vocoder degradation occurred due to noise at the spectrum analyzer input than due to noise applied to the pitch detector. For among the many types of DRT stimuli generated by Gold and Tierney was one in which noise was added to the input to the pitch tracker while the spectrum analyzer was presented with clear speech, and one in which an essentially equal level of noise was added to the spectrum while the pitch detector was presented with clear speech. The resulting DRT scores for words generated in these two conditions revealed that the addition of noise to the input to the pitch tracker degraded intelligibility far less than did the addition of noise to the input to the spectrum analyzer. Indeed, in the former condition, DRT scores averaged 93.0%; in the latter, they averaged 77.2%.

Yet it is possible that the context-free closed-response format of the DRT resulted in "artificially" high scores in the Gold and Tierney study. That is, because of its forced-choice design, the DRT may have been relatively insensitive to the actual difficulty subjects had in labelling stimuli heard in noise (specifically, stimuli with noise in the pitch track). As Pisoni and Koen have noted, such a test design results in higher scores than does one with an open free-response format.[4] Furthermore, the DRT cannot reveal problems in intelligibility beyond the phonological level — problems, e.g., at the syntactic and/or semantic level.

So, in light of previous work conducted on speech intelligibility, and in view of the results of recent work on the intelligibility of vocoded speech, we were led naturally to the following two questions: First, in a test which is more complicated than the DRT (one which places greater cognitive/memory demands upon subjects), how seriously is intelligibility degraded when speech stimuli are vocoded, and what specific types of errors are made in various non-vocoded and vocoded conditions? Second, is there a correlation between subjects' performance scores on the DRT and on a more complex test of speech intelligibility? In order to answer these questions, we designed an experiment using natural and vocoded semantically anomalous (grammatical but meaningless) sentences presented in noise and non-noise conditions.

We used semantically anomalous stimuli in order to strike a balance between subjects' reliance upon acoustic/phonetic cues and their use of syntactic and semantic contextual information. That is, semantically anomalous sentences compel a listener to attend to the acoustic and phonetic information in the signal. As Pisoni and Hunnicutt[3] state, "[Such] sentences permit a somewhat finer assessment of the availability and quality of the 'bottom-up' or acoustic-phonetic information in the stimulus input" (p. 573). Yet because such stimuli are sentential, they also place some of the same memory demands upon listeners as do non-anomalous sentences.

It should be stressed here that, although the present experiment is referred to throughout as a test of sentence intelligibility, it is also a test of sentence recall. Although any intelligibility test necessarily requires recall, the degree to which memory is implicated and the nature of the

memory processes involved vary depending upon the demands of the test. In fact, in suggesting reasons for the decreased intelligibility of synthetic speech, Luce, Feustel, and Pisoni[6] make a distinction between short- and long-term memory. That is, they indicate that short-term memory involves the encoding and/or rehearsal of the perceived signal. Further, they suggest that the perception of synthetic speech places greater demands on short-term memory than does the perception of natural speech, due to the "degraded or impoverished [phonetic] representations" characteristic of synthetic speech (p. 18). The increased demands on short-term memory consequently render the "successful transfer of items from short- to long-term memory" relatively difficult (p. 28). This is an important point since, they suggest, this transfer is essential in a recall task.

A distinction between encoding and recall may thus prove relevant to our analysis of errors made in the four test conditions and in our comparison of these errors with those made on DRTs. It should be kept in mind that some tests may be more difficult than others because they tap different psycholinguistic and cognitive abilities, and these differences may relate to or depend upon encoding and memory processes.

II. EXPERIMENT

A. Subjects

Four groups of seven subjects took part in the experiment, with each group being exposed to one of four test conditions. Twenty-six of the 28 subjects were employees at the M.I.T. Lincoln Laboratory while two were graduate students in the Linguistics Department at Brown University. Six of the subjects were female; 22 were male. The mean age of the subjects in two of the groups was 37; the means for the other two groups were 38 and 35. Overall, the subjects' mean age was 37, with a range of 20 to 51 years. In order to maintain some similarity in the composition of the four groups, we placed at least one female in each group. We also kept the range in age roughly constant across all groups.

Although subjects had had varying degrees of exposure to vocoded speech — ranging from none to extensive — in no group was there a concentration of individuals with a similar degree of exposure. However, in order to be included in the experiment, an individual had to prove capable of performing the task required. Because we did not know what level of performance to expect on the test, we did not establish a performance criterion *a priori*. Rather, we determined whether an individual was to be included after we tabulated the total number of errors he or she made on the test. As a result, we excluded three individuals. (These three were not among the 28 subjects for whom results are reported below.) Two of these three were excluded because they had great difficulty in keeping up with the pace of the test and left scores of items — even entire sentences — blank. The third was not included because, subsequent to taking the test, he revealed that he had a mild reading disability. (This disability was apparent in the nature of the responses he provided and actually led us to question him about his reading and writing skills.)

B. Stimuli

Stimuli in all test conditions consisted of a set of 57 semantically anomalous sentences. In order to construct these sentences, we first compiled a set of relatively common nouns, adjectives, and verbs in which, in word-initial position, each one of 19 consonant phonemes appeared 15 times — six times in nouns, six times in adjectives, and three times in verbs. The phonemes were /p,b,t,d,k,g,č,ǰ,š,f,v,s,z,m,n,l,r,θ,h/. (No attempt was made to balance phoneme frequency in any position other than word initially.) We then selected, at random, words from each of the three part-of-speech categories and inserted them into their appropriate syntactic position, in accordance with the subject-verb-object pattern of the sentences. After completing this procedure, we added the determiners "a" and "the." All sentences exhibited the same syntactic structure such that S →NP + VP where NP → (det.) + adj. + noun and VP → verb + det. + adj. + noun. Thus, each sentence contained either six or seven words. (Sixteen of the sentences had no determiner in sentence-initial position.) There were 383 words in the list of 57 sentences — 285 content words and 98 determiners. Of these, 140 words were monosyllabic and 145 were bisyllabic. No content word occurred more than once although, due to an occasional paucity of phonologically appropriate words, such as adjectives beginning with /z/, it was sometimes necessary to use related forms (e.g., "zing" and "zingy").

The content words varied with respect to the likelihood of their being confused with other words. For example, it was predicted that, while "Mick" might be heard as "Nick" or "mitt" and "tickle" might be heard as "pickle" or "fickle," words such as "bargain" and "luscious" would rarely, if ever, be misunderstood due to the fact that such words cannot occur as members of any minimal pair (e.g., *"bargaim" and *"ruscious"). In this respect, the corpus of test words was similar in composition to words which occur in everyday speech.

Stimuli were presented in each of four test conditions: (1) non-vocoded with no noise added; (2) vocoded with no noise added; (3) vocoded with noise added to the input of the pitch tracker; and (4) vocoded with noise added to the input of the spectrum analyzer. Hereafter these four test conditions will be referred to as CLEAR, VOC, VOCP, and VOCS, respectively.

A tape recording of the 57 sentences was made by one of the experimenters (MM) who read the sentences in a natural manner. These sentences were then vocoded with a real-time channel vocoder (see Gold and Tierney[5] for program description). The Lincoln Digital Signal Processors (LDSPs) — simple programmable computers of a Harvard architecture — were used to implement the real-time channel vocoder program. Noise was generated within the computer so that it could be added to the pitch track without affecting the spectrum, and vice versa. In the VOCP and VOCS conditions, the signal-to-noise ratio at the input was approximately 0 dB. The important fact to emphasize is that identical vocoding programs were used to generate the Gold and Tierney DRT stimuli[5] (labelled "clearpitch" and "clearspec," Gold and Tierney report, p. 15) and the semantically anomalous sentences in noise. Hence, the acoustic properties of the DRT stimuli and the sentences were, in all test conditions, identical.

After being generated all sentences were then recorded on reel-to-reel magnetic tape for later presentation. There was a 20-second interval between the onset of each sentence in all conditions. Each test lasted approximately 20 minutes.

C. Procedure

Twenty-six of the subjects were tested individually or in small groups in a sound-attenuated room at Lincoln Laboratory. Two others (the linguistics graduate students) were tested in a sound-attenuated room in the Linguistics Laboratory at Brown University. Subjects at Lincoln Laboratory heard stimuli played on a TEAC A2340SX tape recorder with a Hewlett Packard 467A amplifier. Subjects at Brown University heard stimuli played on a Technics 1500 tape recorder with a Shure Solo-phone amplifier. All subjects wore AKG headphones, and for all the amplitude was set at a comfortable listening level.

Prior to the test, subjects were told that they would hear grammatical but nonsensical sentences, and they were instructed to write each sentence as soon as they heard it. They were also given an example of a semantically anomalous sentence to ensure that they understood what the nature of the stimuli would be. They were told to guess if they were uncertain or, if they missed a word entirely, to leave a blank.

D. Error Analysis

We analyzed the data in 5 basic ways. These consisted of (1) a general tabulation of **all** errors; (2) a classification of each error as one of several types of "linguistic" errors; (3) a categorization of each error as one of a number of "transpositional" errors; (4) a categorization of each error in terms of the "position-in-sentence" of the word in which the error occurred; and (5) a comparison of overall accuracy scores on the semantically anomalous test and the DRT. The specific details of the categorization and analysis procedures are described below.

1. Overall Error Analysis

We carried out this analysis by counting **every** error which occurred in every word. It was often the case that a single word exhibited more than one error, as in "thickened" → "sicken." Here, there are two errors in the subject's response to the stimulus word, "thickened." These are /θ/ → /s/ and {-ed} → φ. If an entire word was erroneous (e.g. "Tim" → "camera"), it was counted as one error.

2. Linguistic-Error Analysis

A linguistic error was an error classified as phonemic, syntactic, semantic or "other." A phonemic error was one in which a phoneme was affected. Such an error most often involved only a single phoneme (e.g., "thimble" → "symbol"; "chief" → "cheap"; "Chuck" → "chalk"), although it occasionally involved a cluster (e.g., "thighs" → "flies"; "lean" → "green"). A syntactic error was classified as one involving the grammatical structure of the sentence (e.g., "liked" → "likes"; "a" → "the"; "the" → "her"). As is apparent, both word-level (morphological) and sentence-level (syntactic) errors fell into the syntactic-error category. A semantic error was classified as one in which the target word and the word provided by the subject bore some relation in meaning to one another (e.g, "first" → "third"; "nabs" → "snares"; "careful" → "cautious"). Errors which did not fall into one of these three categories were classified as "other." Such errors included metathesis (the transposition of a sound or sounds as in "shines a safe" → "saves a shine") and perseveration (the repetition of a word presented in a sentence immediately preceding the one to be recalled). The category "other" also included errors whose cause could not reasonably be inferred (e.g, "Bob" → "fall"; "Tim" → "camera"; "fake" → "big"). Of course, within a single word, more than one type of error might appear. For example, one subject rendered "chips" as "gypped." In this instance, the response contained both a phonological error in the initial consonant and a syntactic error in the tense of the verb; hence, two errors were counted.

3. Transpositional Error Analysis

A transpositional error was one of substitution, omission, or insertion. A substitution error involved the confusion of one phoneme, morpheme, or word for another (e.g., "thimble" → "symbol"; "liked" → "likes"; "first" → "third". An omission error arose when a phoneme, morpheme or word was omitted (e.g., "master" → "aster"; "newer" → "new"; "chief" → φ). And an

insertion error occurred when a phoneme, morpheme, or word was inserted (e.g., "raid" → "grade"; "jewel" → "jewels"; ϕ → "through"). Clearly, all transpositional errors could be cross classified with linguistic errors.

4. Position-in-Sentence Error Analysis

For this error analysis, we counted all words containing at least one error and classified them with respect to their position in the sentences in which they occurred. Thus, if a subject rendered the anomalous sentence, "A paper nature seeks the cool master," as "The papered nature seeps the cool aster," then there were errors in words 1, 2, 4, and 7 — which also happened to be a determiner, adjective, verb, and noun.

5. Comparison of Semantically Anomalous Test Results With DRT Scores

We compared DRT scores with the percent correct on the anomalous sentence test in each of the four conditions. In this case, percent correct was defined as the average percent of correct words (words in which no error occurred), excluding insertion errors. We also compared the average number of all errors on the sentence test and the DRT. And, in order to evaluate the relative degree of accuracy on both tests, we calculated the ratio of errors in each vocoded condition to errors in the CLEAR condition.

Statistical analyses of the data were conducted with the repeated-measures analysis of variance (ANOVA) (Reference 7) and the Tukey HSD post-hoc test.[8] The primary statistical test was a two-way ANOVA with a $4 \times 3 \times 3$ design (test condition $\times$ linguistic error $\times$ transpositional error). Data from the category "other" were not included in this statistical analysis. The other statistical test was a one-way ANOVA with a 4×7 design (test condition $\times$ word-in-sentence error).

E. Results

The data provided by the experiment were abundant and complex. For the sake of clarity, we have divided discussion of the results into the same five sections described above in the error analysis. Moreover, due to the great preponderance of phonemic substitution errors, we have added a sixth section — one devoted solely to a detailed analysis of the type of phonemic substitutions observed.

1. Results: Overall Error Analysis

Results of our analysis revealed large differences in the number of errors for each of the four test conditions (see Table I and Figure 1). The lowest average number of errors per subject was associated with CLEAR, with an average of 10, and the highest with VOCS, with an average of 141. VOC and VOCP fell between, with averages of 25 and 47, respectively.

TABLE I Total Errors				
	CLEAR	VOC	VOCP	VOCS
Total	70	173	328	989
Average	10	25	47	141
Range	4-15	12-39	39-61	100-194
S.D.	5	12	7	30

Statistical analysis of the total errors revealed a highly significant main effect for condition [$F(3,24) = 85.62$, $p < 0.001$]. A Tukey HSD post-hoc test [HSD (0.05) = 2.584] revealed no significant difference between the means of CLEAR and VOC nor between the means of VOC and VOCP. However, there was a significant difference between the means of CLEAR and VOCP and between the means of VOCS and all other conditions.

2. Results: Linguistic Error Analysis

The number of errors in each of the four linguistic categories — phonemic, syntactic, semantic, and other — are shown in Table II. (Recall that, for the purposes of statistical analysis, the category "other" was not included, although totals for this category are presented in the table.) There was a significant main effect for linguistic error type [$F(2,48) = 85.29$, $p < 0.001$], with total errors in the phonemic, syntactic, and semantic categories being significantly different from one another. There was also a significant test condition $\times$ linguistic-error type interaction [$F(6,48) = 33.73$, $p < 0.001$]. The largest number of errors fell into two VOCS linguistic categories. These were the phonemic and syntactic categories which, in VOCS, had 613 and 239 errors, respectively. Post-hoc analysis [HSD (0.05) = 4.893] revealed that the mean number of VOCS phonemic errors was significantly greater than the mean number of errors in all other nine condition $\times$ linguistic-error categories. The mean number of VOCS syntactic errors was also significantly greater than the mean number of errors in all other condition $\times$ linguistic-error categories with only one exception — the VOCP phonemic category. Further, in both VOCP and VOCS, there were significantly more phonemic than syntactic or semantic errors and in VOCS there were significantly more syntactic than semantic errors. Within both CLEAR and VOC, the mean number of phonemic, syntactic, and semantic errors did not differ significantly.

In spite of the large differences in the totals of errors within the various linguistic categories, the relative proportions of errors in all three vocoded conditions were quite similar (see Table II). That is, the percentage of phonemic errors in VOC, VOCP, and VOCS was about 60%; the percentage of syntactic errors was between about 24% and 32%; and the percentage of semantic

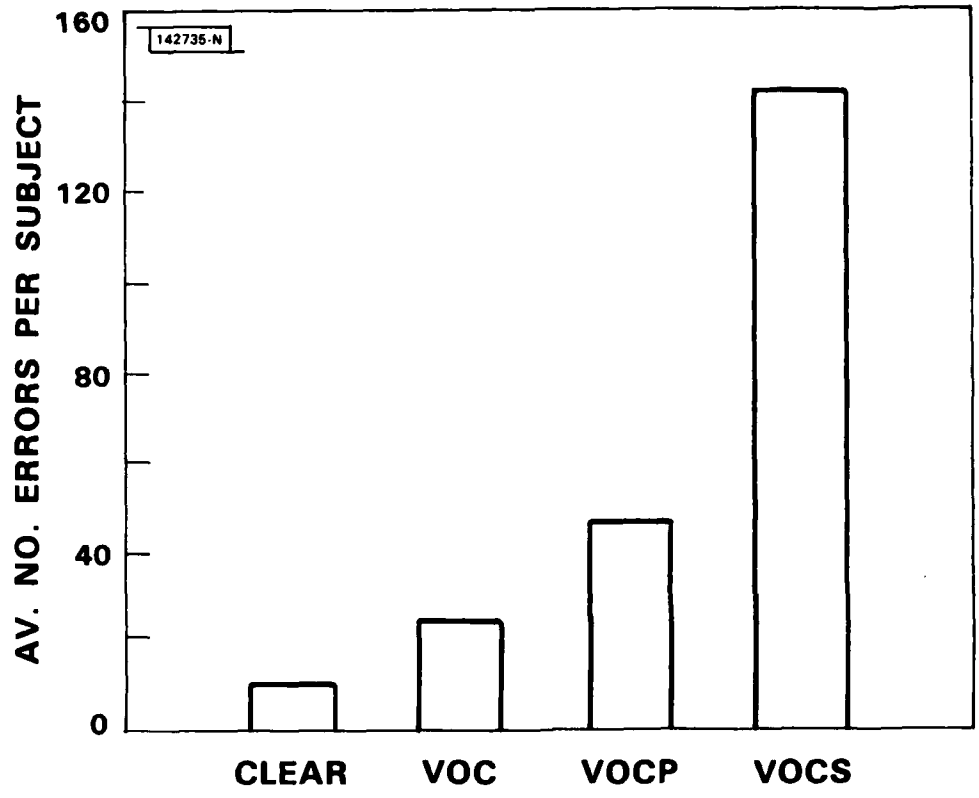


Figure 1. The average number of errors per subject in the semantically anomalous sentence test, in response to four conditions — CLEAR (non-vocoded, no noise), VOC (vocoded, no noise), VOCP (vocoded, noise added to the input to the pitch tracker), VOCS (vocoded, noise added to the input to the spectrum analyzer).

TABLE II

Errors by Linguistic Category: Totals and Percentages*

	CLEAR	VOC	VOCP	VOCS
Phonemic	14 (20.00)	103 (59.54)	186 (56.71)	613 (61.98)
Syntactic	27 (38.57)	45 (26.01)	104 (31.71)	239 (24.17)
Semantic	17 (24.29)	17 (9.83)	33 (10.06)	71 (7.18)
Other	12 (17.14)	8 (4.62)	5 (1.52)	66 (6.67)

*In this and in the following table, percentages are presented in parentheses below each total. Percentages in each column add up to 100.

errors was between about 7% and 10%. (Because subjects made very few errors in CLEAR, and because the mean number of linguistic errors in CLEAR did not differ significantly from one another, it is difficult to draw conclusions regarding the distribution of errors in this condition, although the distribution seems to be quite unlike that of the vocoded conditions. Note, for example, that phonemic errors comprised only 20% of the errors in CLEAR.)

3. Results: Transpositional Error Analysis

The total number of transpositional errors — i.e., errors of substitution, omission, and insertion — are presented in Table III. As is clear, in all four conditions, the largest number of errors were substitutions and the fewest were insertions. There was a significant main effect for transpositional error [$F(2,48) = 239.37$, $p < 0.001$], and a significant cond' 'on $\times$ transpositional-error interaction [$F(6,48) = 64.55$, $p < 0.001$]. The mean number of substitution errors was significantly greater than the mean number of omission or insertion errors in the three vocoded conditions, but not in CLEAR [HSD (0.05) = 1.488]. There was no significant difference between the mean number of omission and insertion errors in any condition.

As was the case with the linguistic errors, the relative proportion of transpositional errors was nearly identical for all three vocoded conditions (see Table III). That is, approximately 79% of all VOC, VOCP, and VOCS errors were substitutions, about 14% were omissions, and 5% to 8% were insertions. (Again, generalizations about the distribution of errors in CLEAR must be

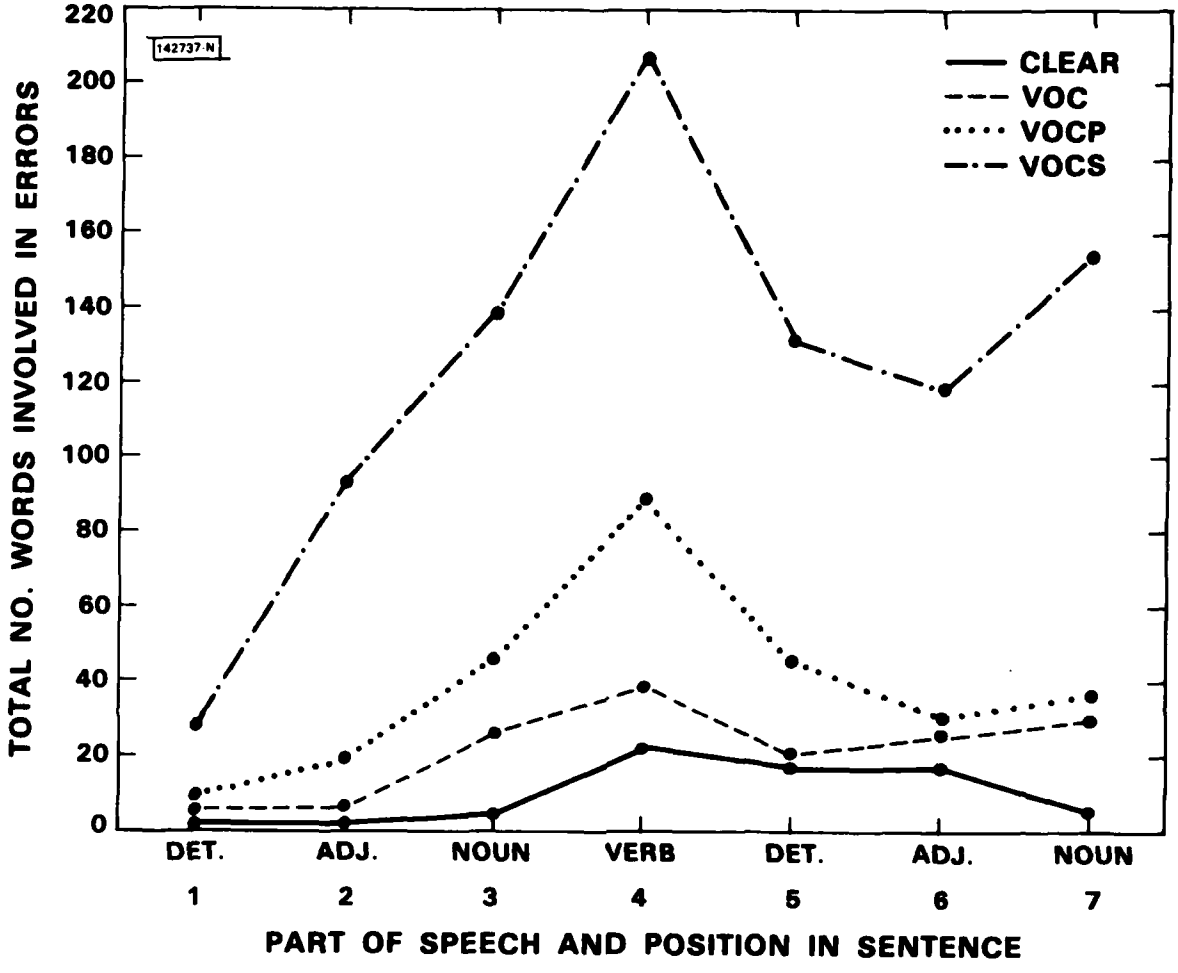


Figure 2. The total number of words involved in errors in all four conditions, plotted according to part of speech and position in sentence.

TABLE III

Errors by Transpositional Category: Totals and Percentages

	CLEAR	VOC	VOCP	VOCS
Substitutions	55 (78.57)	137 (79.19)	260 (79.27)	776 (78.46)
Omissions	13 (18.57)	27 (15.61)	45 (13.72)	131 (13.25)
Insertions	2 (2.86)	9 (5.20)	23 (7.01)	82 (8.29)

made with caution, given that there were few overall errors CLEAR, and given that there were no significant differences in the mean number of substitutions, omissions, and insertions in this category.)

Finally, there was a significant linguistic-error × transpositional-error interaction [$F(4,96) = 71.12$, $p < 0.001$]. Not surprisingly, there proved to be significantly more phonemic substitutions than any other error type [HSD (0.05) = 4.112]. There was also a significant condition × linguistic-error × transpositional-error interaction [$F(12,96) = 19.55$, $p < 0.001$]. Post-hoc analysis revealed that the mean number of phonemic substitutions exceeded all other error types within each vocoded condition [HSD (0.05) = 8.555] as did the mean number of syntactic substitutions within VOCP and VOCS. Moreover, the mean number of phonemic substitutions was significantly greater in VOCS than in VOCP, and in VOCP than in VOC. The mean number of syntactic substitutions was significantly greater in VOCS than in VOCP.

4. Results: Position-in-Sentence Error Analysis

With respect to the position-in-sentence error analysis, there emerged a significant main effect for test condition [$F(3,24) = 22.100$, $p < 0.001$], with far more errors occurring in VOCS than in any other condition (see Figure 2). There was also a highly significant main effect for word-in-sentence error [$F(6,144) = 46.07$, $p < 0.001$], as well as a significant condition × word-in-sentence interaction [$F(18,144) = 9.95$, $p < 0.001$]. Post-hoc analysis revealed that the sentence-medial word (the verb, in position 4) exhibited significantly more errors than did any other word, while the word in sentence-initial position exhibited significantly fewer than did any other [HSD (0.05) = 2.146]. Post-hoc analysis also revealed that there were significant differences between conditions in which stimuli had no added noise (CLEAR and VOC) and conditions in which stimuli did have added noise (VOCP and VOCS). That is, there were no significant word-in-sentence positional effects in CLEAR, and there were only two in VOC. (In VOC, the number of errors associated with the sentence-medial verb in position 4 was significantly greater than the

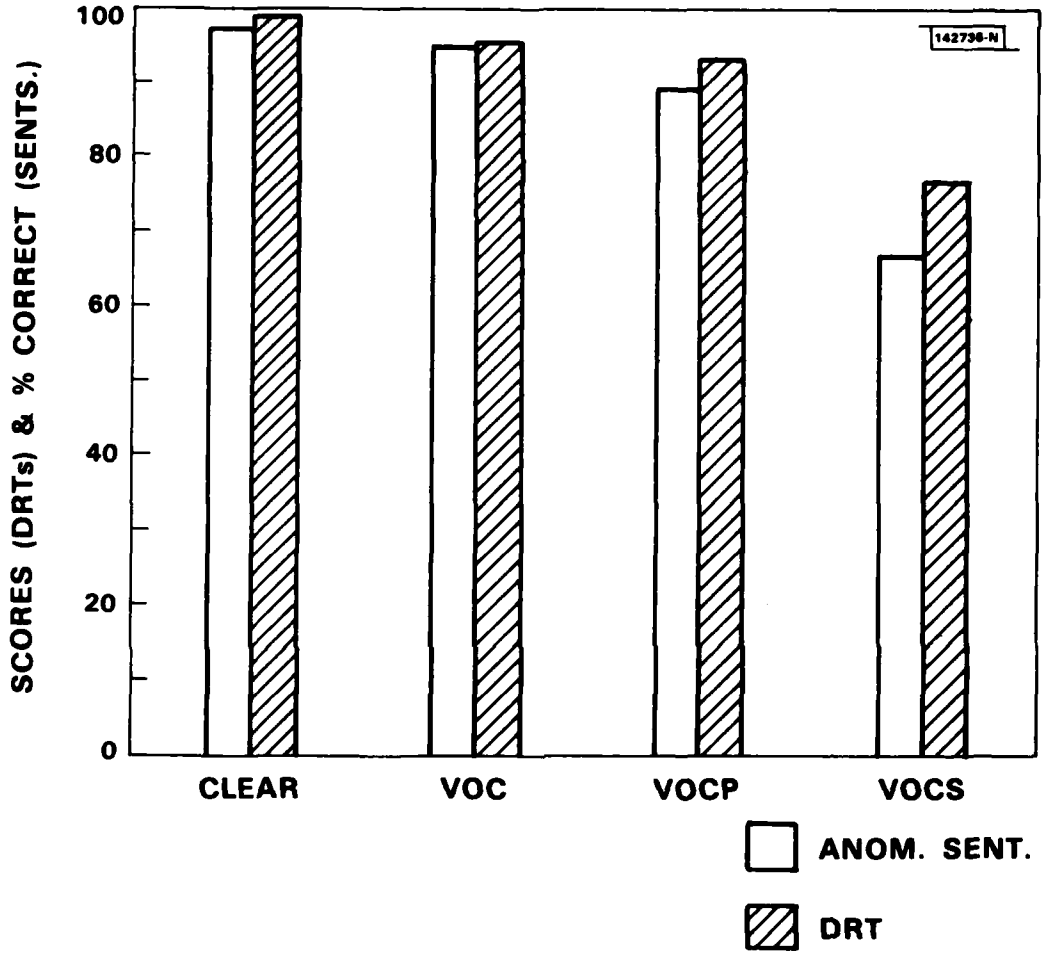


Figure 3. Average DRT scores and anomalous sentence test percent correct in all four conditions. Note the marked similarity in the patterns for the two test types.

number of errors associated with the determiner and adjective in positions 1 and 2.) Yet, in VOCP and VOCS, robust positional effects emerged. In both of these conditions, the word in position 4 had significantly more errors than did any other word. Furthermore, in VOCS, there were significant differences between errors in all positions with only one exception: There was no difference between the number of errors for words in positions 5 and 6 — the second determiner and adjective.

Another observation made as a result of this analysis was that there was a very large difference in the number of errors associated with the first and second determiner (the words in positions 1 and 5) in VOCP and VOCS. Some of this difference can no doubt be attributed to the fact that only 41 of the sentences had a determiner in position 1 while all 57 had a determiner in position 5. Thus, there were 1.39 as many determiners in position 5 as in position 1. Yet examination of the error rates associated with the determiners in these two positions reveals that the difference in frequency of occurrence cannot account entirely for the disparity in errors. For in VOCP and VOCS, there were about five times as many determiner errors in position 5 as in position 1.

5. Comparison of Results with DRT Scores

The order of response accuracy, proceeding from most to least accurate, was CLEAR, VOC, VOCP, VOCS in both tests (see Figure 3). Further, the DRT scores and the percent correct in the sentence test were merely one percentage point apart (98.4% and 97.3%) in the CLEAR condition, while the scores for VOC were essentially identical (94.6% and 94.4%). The scores on VOCP and VOCS were less similar, with an approximately 3-point difference between the two tests on VOCP (93.0% and 89.6%), and a 10-point difference on VOCS (77.2% and 67.0%).

In Table IV, the average number of errors for the DRT and for three sets of errors (overall errors, overall phonemic errors, phonemic substitution errors) in the anomalous sentence test are

TABLE IV

Average Number of Errors:
DRTs and Anomalous Sentence Test

	CLEAR	VOC	VOCP	VOCS
DRT	2	5	7	22
Sentences	10	25	47	141
Phonemic Errors Only	2	15	27	85
Phonemic Substitution Errors Only	2	14	23	71

provided. There was a perfect rank-order correlation between the DRT scores and the sentence-test scores — regardless of the error classification (overall, phonemic, or phonemic substitution). However, there were approximately three to six times as many errors on the sentence test as on the DRT in the vocoded conditions.

In terms of the relative distribution of overall errors in the two tests, the anomalous sentence test and DRT were nearly identical (see Table V). Specifically, in both tests, VOC resulted in two to three times as many errors as CLEAR, VOCP in about four times as many, and VOCS in about fourteen times as many. But, when just phonemic errors were considered (either all phonemic errors or phonemic substitutions only), the ratios for the DRT and sentence tests were quite dissimilar. That is, there were proportionately more phonemic errors made on the sentence test than on the DRT. (Note, for example, that while there were about fourteen times as many errors in VOCS as in CLEAR on the DRT, there were nearly 44 times as many errors in VOCS as in CLEAR on the sentences.)

TABLE V

Error Ratios:
DRTs and Anomalous Sentence Test

	VOC/CLEAR	VOCP/CLEAR	VOCS/CLEAR
DRT	3.5	4.5	14.6
Sentences	2.5	4.7	14.1
Phonemic Errors Only	7.4	13.3	43.8
Phonemic Substitution Errors Only	8.1	13.6	41.5

6. Results: Phonemic Substitution Error Analysis

Because the majority of errors made in the vocoded conditions were phonemic substitution errors, we conducted a detailed analysis of this category. In addition, because the syllabic structure of all content words in the test list was alike only to the extent that all contained a word-initial consonant and a vowel following this consonant, our detailed analysis of phonemic substitutions was devoted solely to these two phonemes.

Confusion matrices for word-initial consonant errors are presented in Tables VIa (VOC), VIb (VOCP), and VIc (VOCS). No matrix appears for CLEAR since there were only four word-initial consonant substitution errors in this condition. In the matrices, there are two more "response" than "target" phonemes since subjects reported hearing two phonemes — /w/ and /y/ — which were not among the word-initial target phonemes. The number of errors associated

TABLE VIa

Vocoded Speech: No Noise Added

Response \ Target	p	b	t	d	k	g	č	j	š	f	v	s	z	m	n	l	r	0	h
p			3		5					1									
b					1						4								
t																			
d																			
k	1																		
g																			
č			1																
j							2												
š																			
f												3							
v		2											1						
s																		13	
z											1								
m															3				
n														7					
l																			
r																			
0												2	2						
h																			
w																			4
y																			
Total	1	2	4	0	6	0	2	0	0	1	5	5	3	7	3	0	0	13	4
%	1.79	3.57	7.14	0	10.71	0	3.57	0	0	1.79	8.93	8.93	5.36	12.50	5.36	0	0	23.21	7.14

TABLE VIb

Vocoded Speech: Noise Added to the Pitch

Response	Target p	b	t	d	k	g	č	j	š	f	v	s	z	m	n	l	r	θ	h
p			1		4														
b	1									1	7							2	
t					1					1									
d																			
k																			1
g																			
č								1											
j																			
š							1												1
f		3			1	1					3	6						4	
v													1						
s										3			1					9	
z																		1	
m											1								
n				1										10					
l																			
r																			
θ			1									2	5						
h			2																
w																			
y																			5
Total	1	3	4	1	6	1	1	1	0	5	11	8	7	10	0	0	0	16	7
%	1.22	3.66	4.88	1.22	7.32	1.22	1.22	1.22	0	6.10	13.41	9.76	8.54	12.20	0	0	0	19.51	8.54

TABLE VIc

Vocoded Speech: Noise Added to the Spectrum

Response	Target p	b	t	d	k	g	č	j	š	f	v	s	z	m	n	l	r	θ	h
p		1	16		1				1									2	2
b	1			2		3				4	6		2						
t	1				4		1			1									2
d		1				3							4				1		
k	6	1	2															2	1
g															1	1			
č			2		1				1									1	
j				3					1				4						
š			1				2			3		1							
f	7	6	3		4						1	9						2	
v		6				1				1			8		1	1		1	
s			5	2		1				6			8					18	
z				3															
m															7	2			
n													1	8					1
l													1		1		2		
r				1							1		7	2	5			1	
θ			2	5							1	2	3						
h	8	1		1														1	
w		2										1	2		2	1	1		
y				1											2				4
Total	23	18	31	18	10	8	3	0	3	15	9	13	40	10	19	5	4	28	10
%	8.61	6.74	11.61	6.74	3.75	3.00	1.12	0	1.12	5.62	3.37	4.87	14.98	3.75	7.12	1.87	1.50	10.49	3.75

with each target phoneme appears at the bottom of the matrix along with percentages. Each percentage indicates what proportion of the total phoneme substitution errors in a given test condition was contributed by a specific target phoneme.

As Table VIa indicates, in VOC, there were 56 word-initial phoneme substitution errors. The most frequently misunderstood phoneme was /θ/, which was heard as /s/ thirteen times and comprised over 23% of the word-initial phoneme substitution errors. The second most frequently misunderstood phoneme was /m/ — heard seven times as /n/ and comprising over 12% of the errors. In VOCP (see Table VIb), there were 82 word-initial phoneme substitution errors. As in VOC, the phoneme /θ/ was frequently misunderstood; it was reported as /s/ nine times and as /f/ four times and it comprised over 19% of the errors. The phoneme /m/ was also frequently misunderstood; it was reported as /n/ ten times and it comprised over 12% of the errors. In addition, /v/ was also often reported incorrectly in VOCP, being heard as /b/ seven times and as /f/ three times. The largest number of word-initial phoneme substitution errors occurred in VOCS (see Table VIc), with a total of 267. As in the other two vocoded conditions, /θ/ was frequently misunderstood. There were 28 errors associated with it, amounting to over 10% of the total. However, in VOCS, both /z/ and /t/ were associated with more errors (40 and 31, respectively) than was /θ/. The phoneme /z/, however, was not mistaken frequently for any single phoneme (unlike /θ/, which was heard 18 times as /s/). Rather, it was confused with ten different phonemes. The phoneme /t/, on the other hand, was primarily confused with only one phoneme — /p/. Phoneme subsitutions were not always bi-directional. In VOCS, for example, /t/ → /p/ 16 times, but /p/ →/t/ only once; and /θ/ → /s/ 18 times, but /s/ → /θ/ only twice.

Overall, most of the frequently confused phonemes differed only in terms of their place of articulation. Errors of voicing (e.g., /t/ → /d/) and of manner (e.g., /b/ → /w/ were extremely rare.

Because no attempt was made to balance the vowels in the word list, an analysis of the phonemic substitutions among post-consonantal vowels was somewhat less straightforward than was an analysis of the phonemic substitutions among word-initial consonants. In fact, because not all vowels occurred as targets an equal number of times, confusion matrices for this class of phonemes have not been provided. A summary of the major vowel substitution errors has been provided instead.

In CLEAR, there was only one vowel substitution error (/a^U/ → /o/); in VOC, there were eight. In the latter condition, the largest number of substitutions involved /ʌ/, heard three times as /a/, and comprising over 37% of the total errors. In VOCP, there were 19 errors. The largest number of errors were associated with the vowel /ɛ/, reported as /I/ four times (over 10% of the total) and with the vowel /ʌ/, reported as /a/ four times. Far more errors emerged in VOCS, which had a total of 36 vowel substitution errors. But, as in VOCP, the vowel /ɛ/ was misunderstood in VOCS relatively often — for it was heard five times as /I/ and three times as /ae/, comprising over 22% of the total errors. The vowel /I/ was also incorrectly reported rather often, comprising over 22% of the errors in this condition.

The most salient result of the phoneme substitution error analysis was the marked difference in error rates for the word-initial consonants and the post-consonantal vowels. There were over six times as many consonant errors as vowel errors with 409 errors in the word-initial consonants and 64 in the post-consonantal vowels. (Recall that "post-consonantal" here refers only to vowels **immediately** following the word-initial consonants.)

III. DISCUSSION

It will be recalled that we carried out the present study in order to address two major issues set forth in the Introduction: (1) to determine the intelligibility of semantically anomalous sentences presented in non-vocoded and vocoded conditions with reference to specific error types and (2) to compare intelligibility scores in the various conditions with previously obtained Diagnostic Rhyme Test scores. Results of our experiment provided us with considerable insight into these and related issues.

A. Spectrum Was Relatively More Important Than Pitch

The most salient result of our experiment involved the degree to which intelligibility suffered when noise was added to the spectrum. In this condition (VOCS), subjects made far more errors than they did in any other. This finding is in agreement with that of Gold and Tierney[5] and provides strong support for the hypothesis that information in the spectrum is far more critical to intelligibility than is information in the pitch track. This finding is also in accord with the observation made by Miller and Nicely[1] who state that, in their experiment, the feature voicing was "discriminable at signal-to-noise ratios as poor as -12 dB whereas the place of articulation [was] hard to distinguish at ratios less than 6 dB, a difference of some 18 dB in efficiency" (p. 349). Of course, place of articulation is based fundamentally upon spectral cues and it serves to distinguish such pairs as /m-n/, /t-k/, and /θ-s/. On the other hand, cues transmitted in the pitch track convey voicing information which serves to distinguish such pairs as /t-d/, /k-g/, and /s-z/. Pitch cues also convey information regarding stress and intonation — the prosodic qualities of speech. While the addition of noise to the pitch track of vocoded speech resulted in a relatively large decrement in intelligibility, this decrement was far smaller than that induced by the addition of noise to the spectrum.

Before proceeding, let us state that, given the manner in which noise was added to the vocoded stimuli, it is not clear whether errors made in response to noisy speech were a result of subjects' misperception of accurately vocoded speech or whether they reflected subjects' "correct" perception of inaccurately vocoded speech. That is, the vocoder may have erred in its speech synthesis due to the presence of noise in its input. An additional possibility is that both the subjects and the computer incorrectly encoded speech in noise. A distinction among the possible sources of error is of interest, but remains beyond the scope of the present project.

What remains important in our analysis and critical for work in speech synthesis is the evidence that spectral cues are more important than pitch-track (fundamental-frequency) cues. One implication of this finding is that it may be possible to specify only general properties of the fundamental frequency and still retain good intelligibility, at least in some conditions.*

---

* A pilot study which we conducted lends support to this notion. In this study, the fundamental frequency (FO) of 140 vocoded sentences was linearly interpolated and noise was added to the output signal. In half of the stimuli, the FO was natural (non-interpolated). In the other half, the FO was interpolated in increments of 20-140 msec. In general, subjects found it very difficult to discriminate between sentences with and without such interpolation, even when they heard sentences having 140-msec segments of interpolation.

For, at the phonetic level, the existence of redundant voicing cues pertaining to first formant cutback, transition length, amplitude of aspiration, and fundamental frequency may render the percept of voicing relatively impervious to voicing-based errors. At the word and sentence level, the existence of syntactic and semantic cues may render the suprasegmental information (e.g., stress and intonation) relatively unimportant. This is not to assert that information carried by the pitch track is negligible. Indeed, subjects who were presented with the VOCP sentences made nearly twice as many errors as did subjects presented with the VOC sentences. Rather, we maintain that, in English, the pitch track is less essential to the accurate perception of speech than is the spectrum.

B. Conditions CLEAR and VOC Were Similar

Another notable result of the present experiment is that, although subjects made more errors in response to non-noisy vocoded speech (VOC) than in response to clear speech (CLEAR), the difference was not significant. This finding suggests that the speech generated by the LDSP vocoder is of very high quality. Given the difficult nature of the response task and the fact that all stimuli were anomalous sentences, such a result seems especially heartening and it too is in agreement with conclusions drawn previously by Gold and Tierney.[5] Still, it must be acknowledged that the relatively small difference between intelligibility scores for CLEAR and VOC could grow larger in a context more demanding or stressful than that provided by the present experiment. For example, the fact that Pisoni[9] observed consistent differences between reaction times for natural and synthetic speech suggests that further comparisons of natural and vocoded speech should be made — perhaps when subjects are required to perform under stress or when they must carry out subsidiary tasks while processing speech.

C. Phonemic and Substitution Errors Predominated

Another of our findings was that there were far more phonemic than syntactic or semantic errors. This attests to the degree to which subjects relied upon acoustic/phonetic cues. In fact, the phonetic percept sometimes overrode syntactic and semantic constraints. As a result, subjects produced ungrammatical sentences such as *"Modern Leslie yield a cheap hat," *"A vain foam denied Vicky liar," and *"The more gold vacates a costly gate." Phonetic cues also occasionally overrode semantic restrictions, as revealed by subjects' production of neologisms such as "kack," "zufter," "loddy," "hoffer," "villous," and "reaves." Such errors attest to the importance of acoustic/phonetic information and suggest that tests designed to evaluate phonological accuracy are appropriate for judging the quality of speech systems, especially if these systems are to be used in environments in which the content of the transmitted speech is not highly predictable. Yet the fact that many syntactic and semantic errors also occurred should not be overlooked. In fact, in the three vocoded conditions, syntactic and semantic errors combined constituted about 40% of all errors. Clearly, when intelligibility is not ideal, the entire linguistic system can be adversely affected.

Just as most errors in the vocoded conditions were phonemic errors, so too were most errors substitutions, rather than insertions or omissions. Thus, subjects tended to correctly perceive that a phoneme, morpheme, or word had been presented, even when they were inaccurate in perceiving or recalling it. In fact, when they made phonemic substitution errors which resulted in neologisms, subjects inevitably adhered to the phonotactic constraints of English. That is, their nonsense words were non-occurrent but possible in English. A related point is that, when subjects made errors — even errors which greatly distorted the phonemic structure of words in the stimulus sentence — they nearly always preserved the syllabic structure of the stimulus. Note, for example, the following responses to two VOCS sentences. One subject reproduced the stimulus sentence, "A shoddy lobby mopped the dense hip," as "A shoddy loddy lanced a dank tip" while another rendered "The thirsty vine finds a giant shop" as "The thirsty vein singed a thirsty shot." In both instances, the phonemic content was greatly altered, yet the syllabic information was preserved.

The findings of our analysis of specific phonemic substitutions revealed generally similar error patterns inhering among the three vocoded conditions. Perhaps most noticeable was that, in all vocoded conditions, /θ/ was involved in at least 5% of the errors, and it was most often confused with /s/. Other frequently confused pairs included /m-n/, /v-b/, /s-f/, and, in VOCS only, /t-p/. In none of the conditions did voicing errors predominate. The nature of the errors lends further support to the claim that the spectrum is more integral to the intelligibility of speech than is the pitch track.

Also important is the fact that some of the phonetic substitutions, such as /t/ → /p/ and /θ/ → /s/, were uni-directional, at least in VOCS. A reasonable explanation for this involves the nature of the acoustic properties of the target sounds and of the added noise. The sound /t/ has more high-frequency energy than does /p/; apparently, the noise added to the spectrum masked the high-frequency energy of the /t/, causing subjects to hear it as /p/. Similar results were obtained by Miller and Nicely[1] in certain of their frequency-response test conditions. (For example, presented with stimuli in a frequency-response range of 200-1200 Hz, their subjects heard /p/ as /t/ 46 times, but they heard /t/ as /p/ 91 times.)

We might wish to invoke this acoustically based hypothesis to account for the fact that our subjects also heard /θ/ as /s/ far more often than they heard /s/ as /θ/. However, if noise added to the spectrum masked the high frequency sounds, and if errors were based solely upon the acoustic properties of the target sounds, then we would expect /s/ to be heard as /θ/ more often, since /s/ has more high-frequency energy than /θ/. Why the reverse pattern emerged may have been due to lexical constraints associated with the target and response words. For when we examine the words used in the sentence test, we discover that the probability of /θ/ → /s/ was actually greater than that of /s/ → /θ/. Of the 15 occurrences of /θ/-initial words, nine could be changed to /s/-initial (real) words (e.g., "thawed" → "sawed," "thin" → "sin," "thighs" → "sighs"). But of the 15 occurrences of /s/-initial words, only two could be changed to /θ/-initial (real) words (i.e., "sane" → "thane" and "sinner" → "thinner"). Obviously, there was an interaction between the acoustic signal and the probability that one phoneme would be substituted for

another. Most likely, the /t/ → /p/ uni-directionality emerged because there was a roughly equal probability that the two phonemes in question would be confused. When the probability was not equal, as in the case of the /θ/ → /s/ confusion, lexical constraints prevailed. Although such an interaction complicates the findings somewhat, it is this very interaction which occurs in the course of speech processing in normal linguistic interaction and it must be taken into consideration.

Perhaps the most revealing finding of the phoneme substitution analysis is that the patterns of error were consistent with those derived by other researchers, notably Miller and Nicely.[1] This is particularly interesting in view of the fact that their stimuli were presented in conditions rather different from ours. Thus certain pairs of phonemes (or certain phonetic features) are inherently more susceptible to error than others — a fact long assumed by speech perception researchers. This is why, in general, similar patterns of errors emerged in VOC, VOCP, and VOCS, in spite of the marked differences in the acoustic properties of the sentence stimuli in these three conditions.

As indicated above, our findings regarding vowel substitution errors were somewhat less conclusive than those regarding consonants, due to the nature of the test stimuli. However, there was a tendency for the vowels /I/, /ɛ/, and /ʌ/ to be involved in more errors than other vowels. In fact, /I/ and /ɛ/ were confused relatively often, as were /ʌ/ and /a/. Generalizations regarding the distribution of vowel errors must be made with caution, since we made no attempt to balance the frequency of the analyzed vowels. Still, our findings agree with the results of the classic experiment conducted by Peterson and Barney[10] on the perception of vowels, and with conclusions drawn by Stevens[11] regarding the quantal nature of certain vowel sounds. Stevens has proposed that certain vowels, such as /i/ and /u/, are more "quantal" (i.e., acoustically stable) and more discriminable than are certain other vowels. This insight should have ramifications for work in synthetic speech. Yet little research has been conducted to determine the nature of vowel errors in vocoded (and noisy) speech. The general assumption has been that vowels are less deserving of study due to their acknowledged perceptual salience (e.g., Reference 12).

D. Errors In Sentence-Medial Words Predominated

Another interesting result was that not all words were affected equally by errors. The sentence-medial word was most frequently associated with errors. There are at least two reasons for this result. It could be due to the fact that all sentence-medial words were verbs inflected for tense and that, as such, they were subject to disproportionately more errors than were other words. Indeed, a frequent source of error in the verbs was the confusion of the present and past tense morphemes, {-s} and {-ed}. It could also be due to the fact that sentence-medial words were **medial**. There is an often observed response effect in experiments of word recall, wherein the first and last items of a list are better remembered than are the medial items.[13-16]

E. Errors in Vocoded Conditions Were Similarly Distributed

Another important finding was that, in spite of the very large differences in absolute error rates obtained in the three vocoded conditions, the relative distribution of errors was nearly identical in all three. It seems that degradation is fundamentally uniform, with the phonological system being the most vulnerable when a signal is not perfectly intelligible. We hypothesize that a different pattern emerges when high-quality natural speech is used. In such a case, errors in the recall of syntactic and semantic information, rather than errors in the perception of phonemes are likely to occur.

F. Scores for DRT and Anomalous Sentences Test Were Nearly Identical

We now turn to our comparison of DRT scores with scores obtained in the present experiment. In spite of the appreciable differences in test design and requirements, both the semantically anomalous sentence test and the DRT revealed an identical pattern of difficulty in the four test conditions, with the clear (natural) speech condition resulting in the fewest number of errors (and in only slightly fewer than in the vocoded non-noise condition) and with the vocoded noise conditions resulting in the largest number of errors (with noise in the spectrum proving considerably more detrimental to intelligibility than noise in the pitch track). Naturally, the anomalous sentence test yielded far more errors than the DRT in absolute terms. For in the sentence test, there was a greater opportunity for subjects to err due to its free-response format. This test also placed far greater demands upon subjects' short-term memory.

However, in terms of the percent of correctly rendered words and the ratios of overall errors made in each vocoded condition to errors made in the natural-speech (CLEAR) condition, the anomalous sentence test and the DRT yielded very similar results. This is somewhat difficult to explain given that the DRT scores and ratios were based only upon errors in phoneme identification while in the anomalous sentence test, the percent correct and the error ratios were based upon phonemic, syntactic, and semantic errors of substitution, omission, and insertion. Hence, it is tempting to conclude that the similarity is due primarily to coincidence. Yet it is also possible that both tests tapped a fundamental and relatively stable mechanism in the speech processing system. This leads us to believe that similar ratios would be found, **regardless** of the nature of the intelligibility task, provided that the linguistic stimuli exhibited the same acoustic characteristics in the test conditions involved. This is in agreement with a conclusion drawn by Williams and Hecker[17] who conducted a study similar to, but less detailed than the present one. They state that "the relation between various test scores is not unique but depends considerably on the type of speech distortion involved" (p. 1005).

Much less similarity emerged when DRT error ratios were compared only with phonemic-error ratios in the sentence test. This was rather surprising. We might expect the error ratios for phonemic substitutions on the sentence test to be quite similar to the ratios on the DRT since there is some comparability between the tests when phonemic substitutions are compared to phoneme identification errors. It leads us to ask if such a finding would emerge if a series of isolated

words were presented in an intelligibility test. That is, do the processing and memory requirements involved in comprehending and reproducing anomalous sentences invariably result in serious decrements in the phonological component — which decrements become proportionately larger as the acoustic quality of the sentences worsens? The answer is, as yet, unclear.

What can be provided regarding our comparative analysis of DRT scores and sentence test results are the following conclusions: In spite of the difference in their design and requirements, both the DRT and the anomalous sentences test resulted in identical patterns of response for all four test conditions. In addition, far more errors were made, in all conditions, on the anomalous-sentence test than on the DRT. Thus it is possible that the DRT underestimates the difficulty subjects have in processing speech — especially speech in noise. It is probably wise to administer several types of intelligibility tests in attempting to determine the absolute quality of synthetic and/or noisy speech. Moreover, because both the DRT and the sentence test resulted in nearly identical percents correct and error ratios, we suggest that there exist certain neurolinguistic and cognitive processes which function consistently in response to specific types of stimuli. For example, high-quality vocoded speech may always result in two to three times as many errors in intelligibility as natural speech, regardless of the test used to evaluate intelligibility.

G. Results Suggest Distinction Between Encoding and Recall Effects

Finally, in the Introduction, we referred to a hypothesis made by Luce *et al.*[6] regarding the encoding of synthetic speech. The crux of their hypothesis was that, because of its "degraded" quality, "synthetic speech may require more processing capacity than does natural speech for maintenance of information in short-term memory and subsequent transfer of information to long-term memory" (p. 18). Their experiment with the recall of natural and synthetic words provides strong support for the notion that the encoding of non-natural speech places greater demands upon certain cognitive processes than does the encoding of natural speech. For their subjects consistently recalled fewer synthetic than natural words in a variety of memory tasks. How does this important finding relate to our experiment?

Although the distinction between encoding and recall may be in need of refining, we assume that — in the present experiment — encoding errors were reflected primarily, although not exclusively, in the phonological component, and that recall errors were reflected primarily, although not exclusively, in the syntactic and semantic components. Our data enable us to determine whether observed decrements in performance were due solely to subjects' poor encoding of the stimuli (whatever its cause), or whether the quality of the stimuli resulted in poor recall of the items presented. There are at least three pieces of evidence which reveal that not all of the errors can be attributed to encoding problems only.

First, in the vocoded conditions, 40% of the errors were not phonemic. This suggests strongly that subjects experienced short-term memory difficulties in reproducing the sentences they heard. For example, such an error as {-s} → {-ed} points to the presence of recall difficulties. One might argue that this error is phonetically based — and thus that it is probably a result of improper encoding. Yet such an explanation is intuitively unsatisfying, for aside from their confusion in these examples, there was virtually no instance of /s/ being heard as [d], [t], or

[əd] — the three pronunciations of the past-tense morpheme {-ed}. Second is the existence of such substitution errors as "thick" → "thin," "careful" → "cautious," and "first" → "third." Although infrequent, such responses reveal that subjects correctly perceived the stimulus words but that, during the process of recall, they transposed them into semantically related forms. And third, there is evidence provided by the analysis of errors associated with the word-in-sentence positions. If we could attribute all errors to problems in encoding, we would find it difficult to account for the fact that nearly five times as many determiners were reportedly incorrectly in position 5 as in position 1.

These findings thus agree with those of Luce *et al.*[6] who conclude that the intelligibility of non-natural speech is both a function of its being poorly encoded and of the additional demands it places on cognitive processes.* Thus, it is possible that, even when synthetic speech seems to be as intelligible as natural speech, it is more difficult to process. In the present experiment, the requirements of the task were clearly made so difficult that problems other than those attributable simply to encoding errors emerged. Although it was not our primary objective to demonstrate the existence of such problems, their emergence has provided us with valuable information regarding the ability of individuals to process and recall speech in various conditions.

* Results of a pilot experiment which we conducted suggest that such an approach is profitable. Using a testing paradigm similar to Rabbitt's,[18] we compared word recall for LPC vocoded speech with recall for natural speech. Subjects who heard the LPC vocoded speech made significantly more recall errors than did those who heard the natural speech.

IV. CONCLUSION

In summary, the results of the four intelligibility tests provided us with a number of interesting results.

One of the most salient results was that intelligibility deteriorated to a much greater extent when noise was added at the input to the spectrum analyzer rather than at the input to the pitch detector. Yet we also observed that intelligibility in vocoded speech with **no** noise added was quite good.

Analysis of the distribution of errors revealed that (1) the most frequent errors were phonemic substitution errors; (2) most errors occurred in sentence-medial words; and (3) a similar pattern of errors emerged in all three vocoded conditions. Furthermore, the results of this experiment were nearly identical to those previously obtained in Diagnostic Rhyme Tests, suggesting that a relatively complex task with an open-ended response format can produce the same **relative** distribution of errors in various listening conditions as does a more simple task with a forced-choice response format.

Finally, it seems likely that, when presented with synthetic or otherwise degraded speech in a test such as the DRT or the one used in the present study, subjects perform less well than they do when presented with natural non-degraded speech because they experience short-term memory difficulty in encoding.

ACKNOWLEDGMENTS

The work reported in this paper was performed at the Massachusetts Institute of Technology Lincoln Laboratory, with the support of the Department of the Air Force under Contract F19628-85-C-0002. We thank Clifford Weinstein for his helpful suggestions on an earlier draft of this paper. We are also grateful to Joseph Tierney for his valuable advice on this project and for his ever-insightful comments.

REFERENCES

1. G.A. Miller, and P.A. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," J. Acoust. Soc. Am. **27**, 338-352 (1954).
2. D.N. Kalikow, K.N. Stevens, and L.L. Elliott, "Development of a Test of Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability," J. Acoust. Soc. Am. **61**, 1337-1351 (1977).
3. D.B. Pisoni, and S. Hunnicutt, "Perceptual Evaluation of MITalk: The MIT Unrestricted Text-to-Speech System," 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing, 572-575 (1980).
4. D.B. Pisoni and E. Koen, "Some Comparisons of Intelligibility of Synthetic and Natural Speech at Different Speech-to-Noise Ratios," J. Acoust. Soc. Am. **71**, S94 (1982).
5. B. Gold, and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," M.I.T. Lincoln Laboratory Technical Report 670 (1983), DTIC AD-A138660/6.
6. P.A. Luce, J.C. Feustel, and D.B. Pisoni, "Capacity Demands in Short-Term Memory for Synthetic and Natural Speech," Human Factors **25**, 17-32 (1983).
7. B.J. Winer, *Statistical Principles in Experimental Design* (McGraw Hill, NY, 2nd ed., 1971).
8. G.A. Ferguson, *Statistical Analysis in Psychology and Education* (McGraw Hill, NY, 5th ed., 1981).
9. D.B. Pisoni, "Speeded Classification of Natural and Synthetic Speech in a Lexical Decision Task," J. Acoust. Soc. Am. **70**, S98 (1981).
10. G.E. Peterson, and H.L. Barney, "Control Methods Used In a Study of the Vowels," J. Acoust. Soc. Am. **24**, 175-184 (1952).

11. K.N. Stevens, "Evidence for Quantal Vowel Articulations," J. Acoust. Soc. Am. **46**, 110A (1969).

12. W.D. Voiers, "Diagnostic Evaluation of Speech Intelligiblity," in *Speech Intelligibility and Speaker Recognition,* ed. by M.E. Hawley (Dowden, Hutchinson and Ross, Stroudsburg, PA, 1977), pp. 374-387.

13. J. Wishner, T.E. Shipley, and M.S. Hurvich, "The Serial-Position Curve as a Function of Organization," Am. J. Psychol. **70**, 258-262 (1957).

14. B.B. Murdock, "The Serial Position Effect of Free Recall," J. Exp. Psychol. **64**, 482-488 (1962).

15. M.Q. Lewis, "Short-Term Memory Items in Repeated Free Recall," J. Verbal Learn. Verbal Beh. **10**, 190-193 (1971).

16. R.G. Crowder, "Audition and Speech Coding in Short-Term Memory: A Tutorial Review," in *Attention and Performance*, ed. by J. Requin (Lawrence Erlbaum, Hillsdale, NJ, 1978), pp. 321-342.

17. C.E. Williams and M.H.L. Hecker, "Relation Between Intelligibility Scores for Four Test Methods and Three Types of Speech Distortion," J. Acoust. Soc. Am. **44**, 1002-1006 (1968).

18. P. Rabbitt, "Recognition: Memory for Words Correctly Heard in Noise," Psychon. Science **6**, 383-384 (1966).

APPENDIX

SEMANTICALLY ANOMALOUS SENTENCES

1. A painted shoulder thawed the misty sill.
2. The bitter seed vexes a valid dinner.
3. The tacky runner judged a short fact.
4. Dingy Doug chips the poor jewel.
5. A golden corner varies the thoughtful keeper.
6. A cotton zebra thickened the chief tickle.
7. The simple rocket picks a new female.
8. A zesty joke gets the nice feather.
9. The shiny shore gives a heavy father.
10. Checkered Sharon gained the chilly hope.
11. Recent Gary sets a messy shower.
12. Fake Chuck finished the hopeful golfer.
13. The vague job savors a jolly garden.
14. A thin jailer checked a meager soap.
15. Moody Tim holds the sane zero.
16. A newer deed shines a safe sinner.
17. A luscious devil helps the good raid.
18. The jealous duster lifted a gaudy cap.
19. The helpful knitter makes a gabby lip.
20. A paper nature seeks the cool master.
21. The bossy vapor shakes a careful victor.
22. Top Jane zapped the tense tot.
23. A dark nail zones the round reason.
24. The kind ladder shoots a dim bed.
25. The gilded nest zipped the dusty tank.
26. The zingy thing liked a late toddler.
27. The soft bargain mixes a thick needle.
28. A shoddy lobby mopped the dense hip.
29. Modern Leslie healed a cheap hat.
30. The charming deck robbed the hot jelly.
31. A jaunty fork raised a vacant cow.
32. The funny heaven reads the shallow pepper.
33. Ready Holly doubts the shabby van.
34. Novel Cathy dipped the loud hopper.
35. A vain foam denies a zippy lime.
36. The third pattern teases a zany tailor.
37. High Mick thanked a zealous chin.
38. Healthy Ned tears the solid rat.

39. Lean Rex takes the pale chowder.
40. A lewd pill leads a pink zing.
41. The bizarre pot needed the best zombie.
42. A partial baker knocked the boring shell.
43. Tipsy Peter keeps the better chopper.
44. The damp vase catches a tiny zeal.
45. A kingly thinker bites a nasty lock.
46. A gorgeous villain chopped the rotten thimble.
47. The southern gift beats the tall thighs.
48. Sure Susan bought a famous thirst.
49. A jagged sailor paid a ripe card.
50. A cheerful thistle pours the fat bean.
51. The zinc mitt carries a lazy basket.
52. A feisty chain fights the fertile money.
53. Vast Bob jabbed a junior pack.
54. The thirsty vine finds a giant shop.
55. The moral gold vacates a costly gate.
56. A normal cheater joined the thorough mess.
57. Rapid Zach nabs a vulgar mirror.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ESD-TR-85-211	2. GOVT ACCESSION NO. AD-A160401	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* The Intelligibility of Non-Vocoded and Vocoded Semantically Anomalous Sentences		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER Technical Report 703
7. AUTHOR(s) Molly A. Mack and Bernard Gold		8. CONTRACT OR GRANT NUMBER(s) F19628-85-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173-0073		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element Nos.33401F and 63735F
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20334		12. REPORT DATE 26 July 1985 13. NUMBER OF PAGES 44
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB, MA 01731		15. SECURITY CLASS. *(of this Report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

DRT	natural speech	semantically anomalous	speech perception
diagnostic rhyme test	noise	sentence	vocoded speech
intelligibility	pitch	spectrum	vocoder

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The present study is devoted to an analysis of the intelligibility of semantically anomalous sentences presented in four acoustically different conditions: (1) natural speech, no noise; (2) vocoded speech, no noise; (3) vocoded speech, noise added to the pitch track; (4) vocoded speech, noise added to the spectrum. One of our objectives was to analyze the specific types of errors in each condition. Our other objective was to compare results of this analysis with results obtained from the Diagnostic Rhyme Test (DRT). Results revealed that intelligibility was quite good in conditions (1) and (2), relatively poor in (3), and quite poor in (4) — results consistent with DRT data. Further, about 60% of all errors were phonemic, while 40% were syntactic and semantic. We conclude that informatin in the spectrum is more critical than information in the pitch track, that most errors affect the phonological component when intelligibility is poor and context is uncertain, and that the DRT is an appropriate though perhaps insufficient test of speech intelligibility.

DD FORM 1 Jan 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

END

FILMED

11-85

DTIC