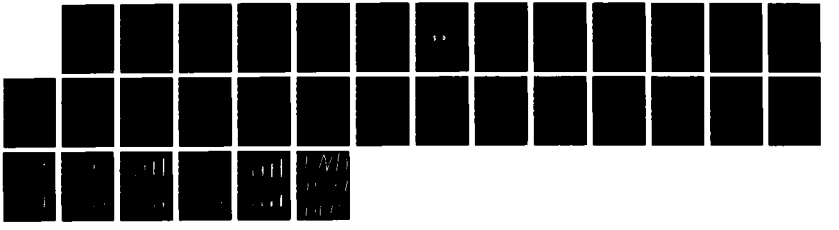
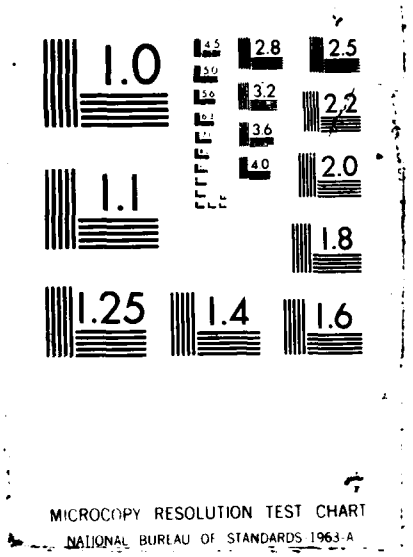
AD-A185 540 ACTIVITIES AS AN ONT POSTDOCTORAL FELLOW(U) NAVAL RESEARCH LAB WASHINGTON DC H J KALLMAN 1987 1/1

UNCLASSIFIED F/G 5/7 NL




1.0
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

(2)

Final Report of Activities as an ONT Postdoctoral Fellow

Howard J. Kallman, Naval Research Laboratory

1987

During the past 12 months, I have been involved in research on the effects of narrowband processing of speech at the Naval Research Laboratory, Washington, D.C. Specifically, the research was designed to address the effects of speech processing on the ability of human listeners to mentally process and understand speech.

Previous research on this issue concentrated largely on intelligibility testing. In particular, the Diagnostic Rhyme Test (DRT) had received widespread use as a method for evaluating the quality of speech after processing. During administration of the DRT, listeners are presented a list of words via headphones that has undergone speech processing of one sort or another (e.g., CVSD, LPC, etc.) As each word is heard, the listener is to choose from among two visually presented words the word that he or she thinks was presented. For example, the word "pond" might be presented and the listener is asked whether the word is "pond" or "bond." Note that in this case, the words differ in terms of whether the first phoneme is voiced or not. If a listener consistently responds correctly to such pairs that include a voicing contrast, the conclusion to be drawn is that the distinctive feature of voicing was preserved by the speech processor. If, on the other hand, many errors are made on items that test the voicing feature, the conclusion is that the processor distorts this feature such that it becomes difficult to perceive. Proper administration of the DRT yields intelligibility scores on six distinctive features of speech as well as an overall score of intelligibility.

The DRT has proven useful as a tool to evaluate the quality of processed speech. However, there are certain limitations to the DRT as an overall test of speech quality. First, the test only evaluates consonants in initial

word position, a limitation that could prove problematic in certain cases. Second, and more significant, is that the words that comprise the test are words that have been spoken in isolation and with careful articulation. There are significant reasons to question whether the results of such a test generalize to the intelligibility of continuous speech because the acoustic events that signify each phoneme may differ for continuous conversational and well-articulated isolated speech. Accordingly, the generalizability of results obtained with the DRT may be somewhat suspect. Third, owing to the nature of the DRT, listeners are encouraged to focus their attention on the surface acoustic characteristics of the presented words; comprehension of the meanings of the presented words is not necessary, just perception of the phonemic information. Related to this last point, the DRT does not test the degree to which suprasegmental information (such as the pitch contour of the sentence or the relative intensities of various syllables and words) is distorted by the processing of the speech. Yet, the presence of accurate suprasegmental information can facilitate the perception and comprehension of words in running speech. Thus although the DRT has shown itself to be a useful tool in evaluating processed speech, limitations are apparent and need to be addressed.

Most of the research that I was involved in at the Naval Research Laboratory used a sentence verification task to provide additional information on the effect of speech processing on listener performance. In a sentence verification task a sentence is presented to the listener--e.g., A robin is a bird--and the listener must indicate whether the sentence is true or false by pressing one of two buttons. Better than chance performance on this task, (chance being 50% correct) requires that the listener understand the words of the sentence along with their meanings. In addition to providing a measure of accuracy in sentence verification,

the task provides a measure of how much additional time it takes to process a sentence that has undergone processing. This is because the listeners are asked to respond as quickly as they can without sacrificing accuracy and these reaction times are recorded. Indeed, it is possible to have a situation in which two processors yield essentially the same intelligibility scores (e.g., very high intelligibility) but more effort is required to comprehend sentences produced by one of the processors, with the result that reaction times are longer.

The first experiment that was conducted during my stay at the Naval Research Laboratory evaluated the effect of bit error rate on sentence verification using speech that had been processed by an LPC-10 processor. A paper based on this research has been submitted to the journal MILITARY PSYCHOLOGY and is attached to this report as Appendix A. Details of the study can be obtained by reading this report and will only be summarized briefly here (an NRL technical report based on this research is also in press).

In the experiment, listeners were presented with sentences for verification that either had not been processed, or had been LPC processed with either 0%, 2%, or 5% random bit errors. One of the main goals was to evaluate the effect of the processing variable on reaction times and errors on the sentence verifcation task. In addition, a manipulation of subject-predicate relatedness was used such that each sentence could be characterized as having either a strong relationship between subject and predicate (A toad has warts) or a weak relationship (A toad has eyes). Previous research had shown that perception of a word in a sentence is often facilitated if the previous words in the sentence create a context that suggests the later word. In the present experiment, our context manipulation was included to evaluate the degree to which context could aid in "filling in" speech information that had been degraded or lost

through processing.

The results showed that reaction times and errors increased with LPC processing and with increasing bit errors. This effect was diminished somewhat when strongly related sentences were presented suggesting that context can compensate in part for the loss of acoustic information brought about by speech processing. For additional details and results, see the appendix.

In a second experiment, we gave listeners the same sentence verification task but provided them with an additional task that they were to perform simultaneously. The task consisted of sorting lotto cards quickly with their non-preferred hand (responses on the sentence verification task were made with their preferred hand). Of interest was whether the same pattern would obtain on the sentence verification task when subjects had to divide their attention between sentence verification and another task. An additional question of interest was whether the rate of card sorting would decrease as the speech processing condition became more difficult. The rationale here was that in addition to making speech perception and comprehension more difficult, processing of the speech might make it more difficult for a listener to perform other tasks simultaneously, a situation that would have serious implications for real-world settings in which speech processors are used.

The results showed that while the average reaction times were approximately 150-200 msec greater overall when the lotto task accompanied the sentence verification task, the general pattern of results on the sentence verification task (e.g., the effect of processor condition, the effect of subject-predicate relatedness, etc.) were as in the first experiment. We are in the midst of exploring whether the 150-200 msec increase in overall reaction time is due to response competition or to the extra cognitive load imposed when two tasks must be performed

simultaneously (some of this research will continue while I am back at SUNY-Albany). Another result that is a bit more straightforward is that the speed of sorting the lotto cards decreased as the speech processing condition became more difficult. Accordingly, it is clear that speech that has been degraded requires more effort to process with the result that there is less attentional capacity available to perform simultaneous tasks. We are currently waiting for the results of some follow-up studies before publishing the just-described experiment; it is anticipated that both an NRL report and a publication will result within 6 - 12 months.

Finally, we conducted an experiment in which listeners were to rate on a continuous rating scale the quality of sentences that were LPC 0%, 2%, or 5% bit error processed, or unprocessed. Three speakers were used to generate the stimulus materials, and for each, three speaking rates were used. The resulting factorial design (4 speech processing conditions X 3 speakers X 3 speaking rates) allows for an assessment of the effect of each variable on the ratings along with the effect of any interactions among variables. The speaking rate variable was of interest because although LPC processing might be expected to interact with speech rate (owing to the fact that the frame size is on the order of 20 msec and the amount of information in a frame would vary with speaking rate), little attention has been paid to this variable. Ten subjects were tested in this experiment but the analysis of the data is at present incomplete. The analysis should be completed within 2 months and depending on the result of the analysis, we may submit a short paper and/or write an NRL report based on the research.

In addition to the experiments reported above, my year was spent learning generally about processed speech and methods for evaluation. All of the above research was conducted in collabaration with my mentor at the Naval Research Laboratory, Dr. Astrid Schmidt-Nielsen.

Appendix A

The Effect of LPC Narrowband Processing and Bit Error Rate on Performance in a Sentence Verification Task

Howard J. Kallman and Astrid Schmidt-Nielsen

Code 5526, Naval Research Laboratory

Running Head: LPC AND SENTENCE VERIFICATION

Abstract

The comprehension of narrowband digital speech with bit errors was tested using a sentence verification task. The difficulty of the verification task was varied by using predicates that were either strongly or weakly related to the subjects (e.g., A toad has warts. / A toad has eyes.). The test conditions included unprocessed speech and speech processed using a 2400 bit/s linear predictive coding (LPC) voice processing algorithm with random bit error rates of 0%, 2%, and 5%. In general, response accuracy decreased and reaction time increased with LPC processing and with increasing bit error rates. Weakly related true sentences and strongly related false sentences were more difficult than their counterparts. Interactions between sentence type and speech processing conditions are discussed.

The Effect of LPC Narrowband Processing and Bit Error Rate on Performance in a Sentence Verification Task

Digital voice transmission methods are becoming increasingly widespread, both for ordinary telephone use and for secure voice communications. At the lower data rates required for many secure voice applications, some loss in speech quality occurs. This can have various consequences for human performance, depending on the severity of the degradation. Even slight losses in quality can lower the scores on intelligibility tests such as the Diagnostic Rhyme Test (DRT), which measures the discriminability of pairs of words differing only in a single distinctive feature (e.g., MOOT-BOOT differs only in nasality). Small losses in intelligiblity may have little effect on whether ordinary speech is comprehended, but greater effort and more time may be required by the listener to understand the speech. With more severe degradations, not only is listener effort further increased, but errors in comprehension will occur. Consequently, in addition to intelligibility tests which measure only errors, it is of interest to investigate methods to assess the time and effort required to comprehend various types of processed speech.

A sentence verification task, in which the listener is required to decide as quickly as possible whether a sentence such as A giraffe has stripes is true or false, can be used to evaluate the amount of time necessary to comprehend simple sentences. To the extent that reaction times are long, it can be assumed that greater processing effort is required to comprehend a particular type of sentence or speech processing condition. Manous, Pisoni, Dedina, and Nusbaum (1985) demonstrated that reaction times on a sentence verification task were longer for synthetic than for natural speech even when all of the words were correctly understood. Pisoni and Dedina (1986) also used a sentence verification task to evaluate the effect of speech processing and found higher error rates and longer reaction times for 2.4 kilobits/sec (kbps) linear predictive coded speech than for wideband speech. Longer reaction times that result from poorer quality speech can have negative consequences for performance. For example, in military combat situations where split-second decisions may be required it may take longer to react appropriately to a degraded speech message, even if the message is correctly comprehended.

For narrowband secure voice communications, a linear predictive coding (LPC) algorithm

operating at 2.4 kbps has been established as the DoD standard (MIL-STD-199-113 or Federal Standard 1015). Owing to the widespread application of this standard, we focused on this type of speech. Versions of this algorithm have been incorporated in the Subscriber Terminal Unit (STU-III) and in the Navy's Advanced Narrowband Digital Voice Terminal (ANDVT) and will consequently be widely deployed. Intelligibility tests indicate that although scores for LPC processed speech are lower than for wideband speech, intelligibility is nevertheless reasonably good, with a score of about 86 on the DRT[1] and 98% correct recognition of the words of the ICAO spelling alphabet and digits (Schmidt-Nielsen, in press). In certain military environments high levels of interference or jamming may occur, which could result in significant decreases in message intelligibility. One way to simulate a high interference transmission situation is to introduce random bit errors into the transmission stream of the LPC processor. For LPC with 5% random bit errors, the DRT score falls to about 75 and only slightly over 90% of the ICAO spelling alphabet words are correctly understood. Although these results and results obtained by the Digital Voice Processor Consortium (Sandy, 1984) suggest that transmissions over LPC systems are reasonably comprehensible in the absence of bit errors, and somewhat less so with increasing bit errors, the effect of LPC processing and bit errors on the amount of time that it takes to respond to a message merits investigation. The present experiment was carried out to evaluate the effect of different levels of digital speech degradation on reaction times and comprehension errors in a sentence verification task.

We were also interested in the effect of context on reaction time to and comprehension of processed speech. Military voice communcations are generally more robust than ordinary communications because they often employ highly distinctive vocabularies that are designed to be intelligible under adverse conditions. In addition, knowledge of the mission context may help to make incoming speech easier to understand, so that accurate communication can be maintained under relatively severe degradations. In other situations, for example, normal conversational speech or high level policy discussions, the communication may be more open-ended and fewer contextual constraints would therefore be available to aid comprehension. Knowledge about how contextual information interacts with the effect of speech processing would be useful when

evaluating a speech processor for use in a particular environment, because it would make it easier to take into account the degree to which context could be used to aid comprehension. We manipulated context in the sentence verification task by using either strong subject-predicate relationships (e.g., Camels have humps) or weak subject-predicate relationships (e.g., Camels have tongues) within the sentences.

The context provided by the early part of the sentence can often be used to help disambiguate later words (e.g., Marslen-Wilson & Welsh, 1978; Salasoo & Pisoni, 1985). Thus, in the sentence Camels have humps, comprehension of the word camels would serve to prime the concept humps, owing to the strong relationship between the two concepts in semantic memory. Accordingly, perception of the sentence should be facilitated and reaction times to verify the sentence should be shorter. In contrast, Camels have tongues expresses a weak subject-predicate relationship, and perception of the word camels would not be likely to facilitate perception of the word tongues. The detrimental effects of LPC processing and bit errors on comprehension should be less for strongly related than for weakly related sentences because the strongly related context should help make the degraded words easier to recognize.

Subject-predicate relatedness should also affect the perception of false sentences, but the overall effect on reaction time should be somewhat different. Although the effect of relatedness may be somewhat smaller for false than for true sentences because the relatedness of the subject and predicate would tend not to be as strong, a relatively highly related context should still help perception more than a weakly related one, owing to the priming effect of the earlier words in the sentence on the later words.

In addition to influencing the perception of the words in the sentence, the subject-predicate relatedness variable can also affect decision time, the time it takes to decide that the sentence is true or false once the words of the sentence have been perceived. Strongly related true sentences express relationships that are more closely associated in semantic memory than weakly related ones and are therefore easier to verify, which results in faster reaction times (e.g., Collins and Quillian, 1969, 1972; Glass, Holyoak, & O'Dell, 1974; Rosch, 1975). This effect would probably not be influenced by the difficulty of the speech processing condition because the decision process would

occur after the words of the sentence had been perceived. However, for false sentences, the decision about whether the content of the sentence is true or false would be more difficult in the strongly related case (Collins and Quillian, 1972; Glass et al., 1974). That is, A fiancee is a relative would generally be more difficult to reject at the decision stage than A fiancee is furniture, because fiancee and relative are associated concepts whereas fiancee and furniture are not. As with true sentences, the effect of subject-predicate relatedness on the decision stage of processing should remain roughly constant across levels of speech degradation because it is due to decision processes that should be relatively unaffected by the quality of the sensory information. False sentences, however, contrast with true sentences in that strong relatedness has a positive effect on word recognition but a negative effect on the decision stage. Thus, as the quality of the sensory information suffers with increasing degradation of the speech signal, the advantage of weakly related sentences in terms of decision processes would be counterweighed by the advantage of strongly related sentences in terms of perceptual processes, and the advantage of the weakly related false sentences would diminish with LPC processing and with increasing bit errors.

Finally, practice with a particular type of speech processing should result in improved listener performance. The present experiment included a comparison of performance during the first and second halves of testing. The variables of interest were, therefore, the speech processing condition, subject-predicate relatedness, and first versus second half of testing. In addition to main effects involving these variables, some interactions of these effects with the truth value of the sentences were predicted.

Method

Test materials

There were 96 true and 96 false sentences, generated so that the subjects and predicates in half of the sentences were strongly related and the subjects and predicates in the other sentences were weakly or not related. The true sentences were generated by drawing on previously published norms and lists of strongly and weakly associated or related property and category relationships (e.g., Battig and Montague, 1969; Lorch, 1981; McCloskey & Glucksberg, 1979; Rosch, 1975), with additional items having similar relationships selected and agreed upon by both authors. The

false sentences were generated analogously by choosing untrue properties and categories that were either strongly or weakly related to the item in question. For example:

	Strong	Weak
True Property:	A toad has warts.	A toad has eyes.
True Category:	A fly is an insect.	A gnat is an insect.
False Property:	Camels have horns.	Camels have chimneys.
False Category:	Crabs are fish.	Redwoods are fish.

Sixty additional sentences were generated similarly for a practice list and for fillers. The practice list and the eight test lists had 28 items each. The first 4 items (2 true and 2 false) in each test list were fillers and were not scored. The remaining 24 items in each list were the test sentences consisting of equal numbers of true and false statements equally distributed across strong and weak property and category relationships. The order of the sentences within each list was randomized. The practice list and the test lists were recorded by a male speaker whose voice was known not to create any unusual problems when processed by the LPC algorithm. Approximately 2 sec of silence separated consecutive sentences.

Voice Conditions

In addition to high quality unprocessed speech, there were three versions of degraded speech: LPC processed speech with 0%, 2%, and 5% random bit errors. The LPC tapes were generated by processing the tape recorded materials through a TRW low data rate voice terminal which uses version 43 of the DoD standard LPC-10 algorithm. For the 2% and 5% bit error conditions, random bit errors (as opposed to "burst" errors) were introduced into the bit stream between the analysis and synthesis portions of the processing.

Design

Four counterbalanced sequences of the eight test lists were prepared. Each sequence was divided into halves, with one test list for each of the four processing conditions in each half. The order of the processing conditions was balanced across sequences, but the order of the eight sentence lists remained the same across sequences, so that each set of sentences occurred under all four processing conditions. To further balance possible effects of practice or fatigue, the order in

which the different processing conditions were presented in the second half of each sequence was the reverse of the order in the first half.

Subjects and Procedure

The listeners were 48 undergraduate psychology students at the University of Maryland (12 for each of the four sequences) who volunteered to participate for extra course credit. The listeners were tested individually, with the speech heard through high quality headphones. Prior to the sentence verification task, the listeners were familiarized with the sound of LPC speech by listening to LPC processed versions of five different speakers each reading the same 30 sec paragraph. During the experiment, the listeners were seated at a table and placed the index and middle fingers of their preferred hand on two push buttons labelled TRUE and FALSE. They were to decide whether each sentence was true or false and to push the appropriate button as quickly as possible while avoiding mistakes. The practice list of 28 sentences consisting of LPC processed speech with 2% bit errors was presented just prior to the test lists. After the practice, each listener heard one of the sequences of eight test lists, with a 5 - 10 minute break between the first and second half of testing.

Scoring Procedure

The responses and reaction times were collected and stored by an IBM PC computer. The reaction times were calculated from the end of the last word of each sentence as determined by visual inspection of the digitized waveform.

Results

Analyses of variance were performed on the reaction time and response error data. In the reaction time analysis only correct responses were included. In the analyses, the within subjects variables were processing condition, truth value, subject-predicate relatedness, and replication. The degrees of freedom for the F tests were corrected, where appropriate, for violations of sphericity using the Huynh and Feldt (1976) correction.

As expected, both mean reaction time and error rate were greater for LPC than for unprocessed speech and increased progressively with increases in bit error rate. Mean reaction time was 330 ms for the unprocessed speech and 448, 516, and 627 ms for LPC with 0%, 2%, and 5%

bit errors, $F(2.42, 113.77) = 101.24$, $p < .001$, $MSe = 59{,}023$. The corresponding error rates were 6.0%, 9.9%, 12.4%, and 21.9%, respectively, $F(3.29, 137.00) = 137.71$, $p < .001$, $MSe = 127.79$.

When averaged across processing conditions, the main effect of subject-predicate relatedness was significant with strongly related sentences having shorter reaction times, $F(1, 47) = 8.90$, $p < .01$, $MSe = 26{,}276$, and fewer errors, $F(1, 47) = 8.57$, $p < .01$, $MSe = 159.66$, than weakly related sentences. There was no advantage of strong relationship for the unprocessed speech, presumably because strong trues but weak falses have the advantage with respect to decision time. The overall effect is mainly the result of the increasing advantage for the strongly related sentences with increasing degradation, as evidenced by the significant interaction between processor and subject-predicate relatedness for reaction times, $F(2.27, 106.77) = 4.64$, $p < .01$, $MSe = 36{,}819$, and errors, $F(2.12, 99.77) = 7.95$, $p < .001$, $MSe = 170.95$, shown in Figure 1. In both instances, the effect of processing condition was greater for weakly than for strongly related sentences, presumably because the more strongly related final word was more likely to have been primed or activated by the preceding portion of the sentence and would therefore be easier to recognize even when the speech was degraded.

Insert Figure 1 about here

Averaged across conditions, true sentences were responded to faster than false sentences, and there were fewer errors for false than for true sentences. Mean reaction times were 404 ms for true and 557 ms for false sentences, $F(1, 47) = 170.42$, $p < .001$, $MSe = 52{,}892$. The respective error rates were 15.5% and 9.6%, $F(1, 47) = 46.71$, $p < .001$, $MSe = 282.30$. At first glance, these results might appear to suggest a speed-accuracy tradeoff. However, it is more likely that the low error rate for false sentences reflected a bias toward responding "false" when the listener could not understand all of the words, because the proportion of false responses also increased as the speech became more degraded.

There were significant interactions between truth value and processing condition for reaction

time and for errors (Figure 2). The more difficult processing conditions led to greater increases in reaction time for false than for true sentences, $F(2.00, 93.77) = 6.04$, $p < .01$, $MSe = 35{,}936$. If it is inherently more difficult to decide that a sentence is false, then it may be that decreasing the intelligibility of the speech interacts to make this decision even harder. The error rates, on the other hand, increased more for true than for false sentences $F(2.85, 133.77) = 39.63$, $p < .001$, $MSe = 159.32$. If the listeners had a bias to respond "false" when they could not understand a sentence properly, this would have had the effect of depressing the number of correct true responses while inflating the number of correct false responses. Moreover, this effect would be expected to increase as the speech becomes progressively less intelligible.

Insert Figure 2 about here

The interactions involving subject-predicate relatedness and truth value were of particular interest. Although it was predicted that responses to true sentences would be faster and more accurate for strongly rather than weakly related sentences, a different set of predictions had been made for false sentences. Strongly related false sentences express relationships that can be difficult to distinguish from true ones. As a result, additional time would be required at the decision stage to respond to strongly related false sentences, even though word recogniton may be facilitated due to priming by the strongly related early part of the sentence. Furthermore, because strongly related false sentences express relationships that are harder to distinguish from similar true ones (some listeners may not know for certain whether or not a camel has horns), the error rates for these sentences should be higher than those predicted on the basis of intelligibility difficulties alone. In support of this proposition, when unprocessed speech was presented, the error rates for weakly related false sentences was 11.8%. Because the error rate for strongly related falses was only 0.5%, it can be assumed that unprocessed speech provides little in the way of intelligibility difficulties, and the difference must be attributed to errors made at the decision stage.

As predicted, reaction times were faster (321 vs. 487 ms) and error rates lower (9% vs. 22%) for strongly than for weakly related true sentences, whereas reaction times were faster (499

vs. 615 ms) and error rates lower (5% vs. 14.2%) for weakly rather than for strongly related false sentences. The relatedness by truth value interactions were significant both for reaction time, $F(1, 47) = 178.56$, $p < .001$, $MSe = 42{,}965$, and for errors, $F(1, 47) = 163.02$, $p < .001$, $MSe = 288.59$. The three way interaction of truth value, subject-predicate relatedness, and processor was not significant for the reaction times, $F < 1$, $MSe = 28{,}325$, but it was for the errors, $F(2.74, 128.79) = 3.59$, $p < .02$, $MSe = 141.11$ (Figure 3). For true sentences, the effect of relatedness increased as the degradation of the speech increased, reflecting the increased value of contextual information as the speech became progressively degraded. In contrast, the advantage of weakly over strongly related false sentences decreased with increases in speech degradation, a result that also reflects the increased value of context as the speech degradation increased.

Insert Figure 3 about here

From the first replication to the second, mean reaction times decreased from 525 ms to 436 ms, $F(1, 47) = 53.39$, $p < .01$, $MSe = 59{,}944$, and replication did not interact with processing condition for reaction time, $F(2.36, 111.05) = 1.11$, $p > .10$, $MSe = 45{,}820$. The mean error rate decreased from 13.7% to 11.4% from the first to the second replication, $F(1, 47) = 9.85$, $p < .01$, $MSe = 214.09$, and the effect of processing condition on errors was smaller for the second than for the first replication, $F(2.66, 125.04) = 4.52$, $p < .01$, $MSe = 168.18$ (Figure 4).

Insert Figure 4 about here

The three way interaction of processor, subject-predicate relatedness, and replication was significant both for the reaction times, $F(2.75, 129.04) = 4.53$, $p < .01$, $MSe = 22{,}019$, and errors, $F(3, 141) = 5.27$, $p < .01$, $MSe = 124.79$ (Figure 5). During the first replication, the effect of processor did not differ for weakly related and strongly related sentences. In contrast, during the second replication, the effect of processor was greater for the weakly related sentences. Apparently, context was used more effectively to overcome speech degradations after listeners had

become relatively practiced on the task.

Insert Figure 5 about here

Relationship to Previous Results

A comparison of the present results to previously obtained DRT and ICAO spelling alphabet scores is shown in Table 1. Although the low number of data points precludes the drawing of strong inferences about the functional relationships between the different measures of speech quality, it nevertheless appears that within the range of tested values, an increase of one point in the DRT results in a decrease in reaction time on the order of 10 - 20 ms in sentence verification, depending on factors such as the level of context, and so forth.

Insert Table 1 about here

Discussion

The present experiment used a sentence verification task to test the comprehension of digitally processed sentences using the DoD standard LPC 2.4 kbps algorithm with and without random bit errors. The processed speech conditions tested here had been previously evaluated using the DRT test and the ICAO spelling alphabet (Sandy, 1984; Schmidt-Nielsen, in press). The current approach, which required that the responses of listeners be based on comprehension of the content of each message, was motivated by the desire to obtain additional information about the effects of LPC processing and bit errors on speech effectiveness in the "real world." The results with the sentence verification task were systematic and interpretable within the framework that we outlined in the introduction.

Not surprisingly, reaction times and errors increased with increases in speech degradation. This was expected based on previously obtained DRT scores, and on Pisoni and Dedina's (1986) finding that LPC speech with no bit errors yielded more errors and longer reaction times on a sentence verification task than wideband speech. The increased errors with LPC speech suggest

that relatively inexperienced users may have some trouble using LPC systems for ordinary, unconstrained conversational speech, although the improvement from the first half to the second half of the present experiment indicates that this difficulty might be reduced by practice. The large numbers of errors for LPC at the 2% and 5% bit error rates suggest that an open-ended vocabulary can be very difficult to understand under conditions of high bit errors. However, previously reported results using the ICAO spelling alphabet suggest that there would be substantially fewer comprehension errors when military or other constrained vocabularies are used.

Even in situations where the comprehension errors for a particular processor condition are at an acceptable level, the added processing time required to understand the speech should also be taken into account in determining its acceptability. Our results suggest that the added time required to comprehend a simple sentence using LPC with 5% bit errors is on the order of 250 - 350 ms over that for an unprocessed sentence. This represents sufficient time for a typical adult to scan about seven digits in short term memory (Sternberg, 1966) or access four or five labels in long term memory (Card, Moran, & Newell, 1983). It is therefore likely that the additional time required to comprehend LPC speech with 5% bit errors would detract from other ongoing cognitive activities, a situation that might prove unacceptable if the listener has to respond quickly or engage in simultaneous tasks. Indeed, for some situations our estimates of the extra time required to comprehend LPC processed speech may be low. Pisoni and Dedina (1986) found that reaction times using LPC at 2.4 kbps with no bit errors were more than 1 second longer than for 16 kbps wideband speech. Pisoni and Dedina's higher values for the additional time needed to comprehend LPC sentences may be the result of the minimal amount of practice they gave their listeners with LPC speech (i.e., exposure to only four sentences). Alternatively, it is possible that their speakers' voices were less suited to LPC processing than ours or that some other factor was responsible for the different results.

With increasingly degraded speech, true sentences that expressed strong subject-predicate relationships were responded to more accurately and quickly than weakly related true sentences, and the effect was greatest for the most severely degraded condition, LPC with 5% bit errors. It is reasonable to expect that the importance of context would increase with increasing degradation of

the speech signal. Given a high quality speech signal, it should be possible to understand all of the words of a well articulated sentence in the absence of contextual information. This contention is supported by the fact that in the unprocessed condition there were almost no errors for the weakly related false sentences, where there is little context to aid the comprehension stage. With unprocessed speech, perception could be based on data driven (or bottom-up) processes because the speech data themselves are, in this case, sufficient to define the stimuli unambigously. In contrast, with degraded speech, the acoustic sensory information might not be sufficient for accurate perception of the words. Accordingly, conceptually driven (or top-down) processes would be required to "fill in" missing stimulus data. Because the critical words in the strongly related sentences would have tended to prime one another, missing stimulus information would have been compensated for by knowledge based on the context. The context of the weakly related sentences would be unlikely to prime or otherwise aid the recognition of words that might not be identifiable solely on the basis of the degraded stimulus information, and as a result, performance should have suffered.

In contrast to our results, Pisoni, Manous, and Dedina (1986), in an experiment that tested the effect of sentence predictability on the perception of synthesized speech, found no interaction of predictability and speech type. The two studies are not directly comparable, however. For example, they used high quality synthetic speech as opposed to processed natural speech, and differences in intelligibility among the various types of tested speech were greater in our study. Furthermore, they manipulated sentence predictability whereas we manipulated subject-predicate relatedness. Whether either of these variables, or some other difference between experimental stimuli or procedures, was responsible for the different patterns of results is a question that can only be answered by future research.

It is well known that narrowband digital speech becomes easier to understand after practice, but we did not know how the improvement in performance due to practice would interact with the ability to use contextual information. As expected, faster reaction times and fewer errors were found during the second half of the experiment than the first. In addition, the advantage of strongly over weakly related sentences in the degraded speech conditions was greater in the second half than

in the first half. It seems likely that while the LPC speech was still relatively novel, listeners needed to devote most of their attention to learning how to listen and that this limited the attentional resources available to make use of contextual information. As the listeners became more familiar with the degraded speech, the mental processing of the speech may have become more automatic and attentional resources could then be freed to use the contextual information for top-down processing. This suggests that even though communicators experienced with LPC systems may perform very well in contexts in which they know what to expect, a novel or unexpected message could lead to errors and/or longer reaction times, especially in situations where the speech is further degraded by bit errors.

References

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut Category norms. Journal of Experimental Psychology Monograph, 80, (3, Pt. 2).

Card, T. P., Moran, T. P., & Newell, A. (1983). The Psychology of Human-Computer Interaction, Hillsdale, NJ: Erlbaum.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247.

Collins, A. M. & Quillian, M. R. (1972). Experiments on Semantic Memory and Language Comprehension. In L. W. Gregg (Ed.) Cognition and Learning in Memory (pp. 117-137). New York: Wiley.

Glass, A. L., Holyoak, K. J., & O'Dell, C. (1974). Production frequency and the verification of quantified statements. Journal of Verbal Learning and Verbal Behavior, 13, 237-254.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and splitplot designs. Journal of Educational Statistics, 1, 69-82.

Lorch, R. F. (1981). Effects of relation strength and semantic overlap on retrieval and comparison processes during sentence verification. Journal of Verbal Learning and Verbal Behavior, 20, 593-610.

Manous, L. M., Pisoni, D. B., Dedina, M. J., & Nusbaum, H. C. (1985). Comprehension of Natural and Synthetic Speech Using a Sentence Verification Task. Research on Speech Perception Progress Report No. 11. Bloomington, IN: Indiana University.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. Cognitive Psychology, 11, 1-37.

Pisoni, D. B., & Dedina, M. J. (1986). Comprehension of Digitally Encoded Natural Speech Using a Sentence Verification Task (SVT): A First Report. Research on Speech Perception

Progress Report No. 12. Bloomington, IN: Indiana University.

Pisoni, D. B., Manous, L. M., & Dedina, M. J. (1986). Comprehension of Natural and Synthetic Speech Using a Sentence Verification Task: II. Effects of Predictability on the Verification of Sentences Controlled for Intelligibility. Research on Speech Perception Progress Report No. 12. Bloomington, IN: Indiana University.

Rosch, E. H. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 104, 192-233.

Salasoo, A., & Pisoni, D. B. (1985). Interaction of knowledge sources in spoken word identification. Memory and Language, 24, 210-231.

Sandy, G. F. (1984). Digital Voice Processor Consortium Final Report MTR-84W00053. McLean, VA, Mitre Corp.

Schmidt-Nielsen, A. (in press). The Effect of Narrowband Digital Processing and Bit Error Rate on the Intelligibility of ICAO Spelling Alphabet Words. IEEE Transactions on Acoustics, Speech, and Signal Processing.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

Author Notes

The authors contributed equally to this article. Portions of this article were presented at the 113th meeting of the Acoustical Society of America in Indianapolos, IN, May, 1987. The research was conducted while Howard J. Kallman held an Office of Naval Technology postdoctoral fellowship. The authors thank Corinne Meijer for her assistance in conducting the listener tests, Larry Fransen for his help with the digital processing, and Don Kallgren for lending us his speaking voice. Correspondence can be addressed either to Astrid Schmidt-Nielsen, Code 5526, Naval Research Laboratory, Washington, D.C. 20375, or to Howard J. Kallman, Department of Psychology, State University of New York at Albany, 1400 Washington Avenue, Albany, NY 12222.

Footnote

[1]The DRT scores reported in this paper are scores obtained using the TRW processor that was used to process the speech samples used in this experiment. This processor employs Version 43 of the DoD standard LPC-10 algorithm. The scores reported by the Digital Voice Processor Consortium (Sandy, 1984) are slightly higher, and preliminary results indicate that the new LPC-10e can be expected to score 3-5 points higher than the DRT scores reported here.

Table 1. Comparison of the results of the present experiment with previously obtained DRT scores and percent correct responses on the ICAO alphabet. Strong and Weak refer to average scores for the strongly related and weakly related sentences.

Processing Condition	DRT Score	ICAO Score	% Correct Strong	%Correct Weak	Mean RT Strong	Mean RT Weak
Unprocessed	97.6	99.0	92.7	95.4	341	318
LPC 0% errors	86.0	98.0	90.6	89.6	442	455
LPC 2% errors	81.9	96.2	89.1	86.1	501	532
LPC 5% errors	75.4	90.3	81.2	75.0	588	667

Figure Captions

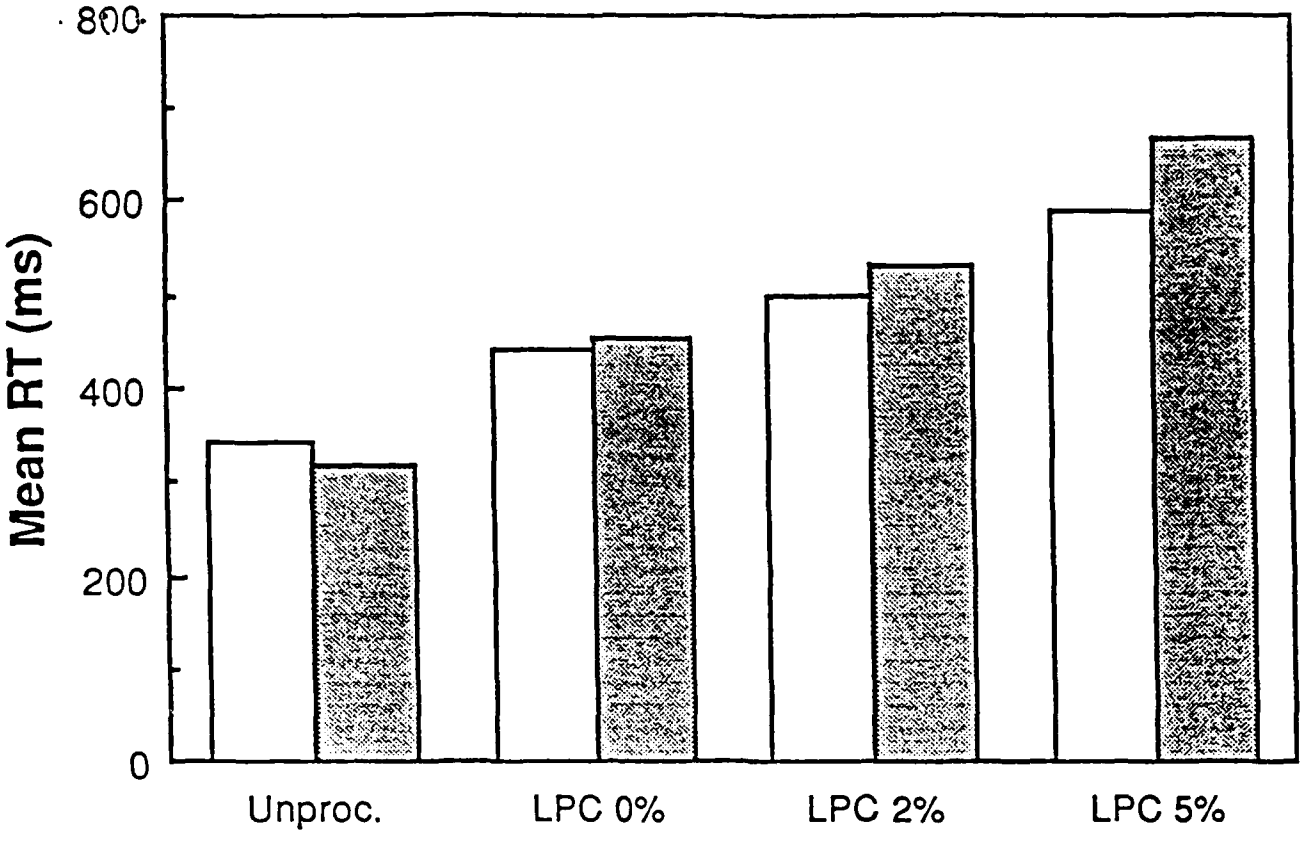
Figure 1. Performance as a function of subject-predicate relatedness (strong vs. weak) and speech processing condition. Mean reaction times (RTs) are shown in the upper panel and mean percentages of errors in the lower panel.

Figure 2. Performance as a function of truth value of the sentence and speech processing condition. Mean RTs are shown in the upper panel and mean percentages of errors in the lower panel.

Figure 3. Performance as a function of subject-predicate relatedness, the truth value of the sentence, and speech processing condition. Mean RTs are shown in the upper panel and mean percentages of errors in the lower panel.

Figure 4. Performance as a function of replication (1st half vs. 2nd half of experiment) and speech processing condition. Mean RTs are shown in the upper panel and mean percentages of errors in the lower panel.

Figure 5. Performance as a function of subject-predicate relatedness, replication, and speech processing condition. Mean RTs are shown in the upper panel and mean percentages of errors in the lower panel.


800
600
400
200
0
Mean RT (ms)
Unproc.
LPC 0%
LPC 2%
LPC 5%

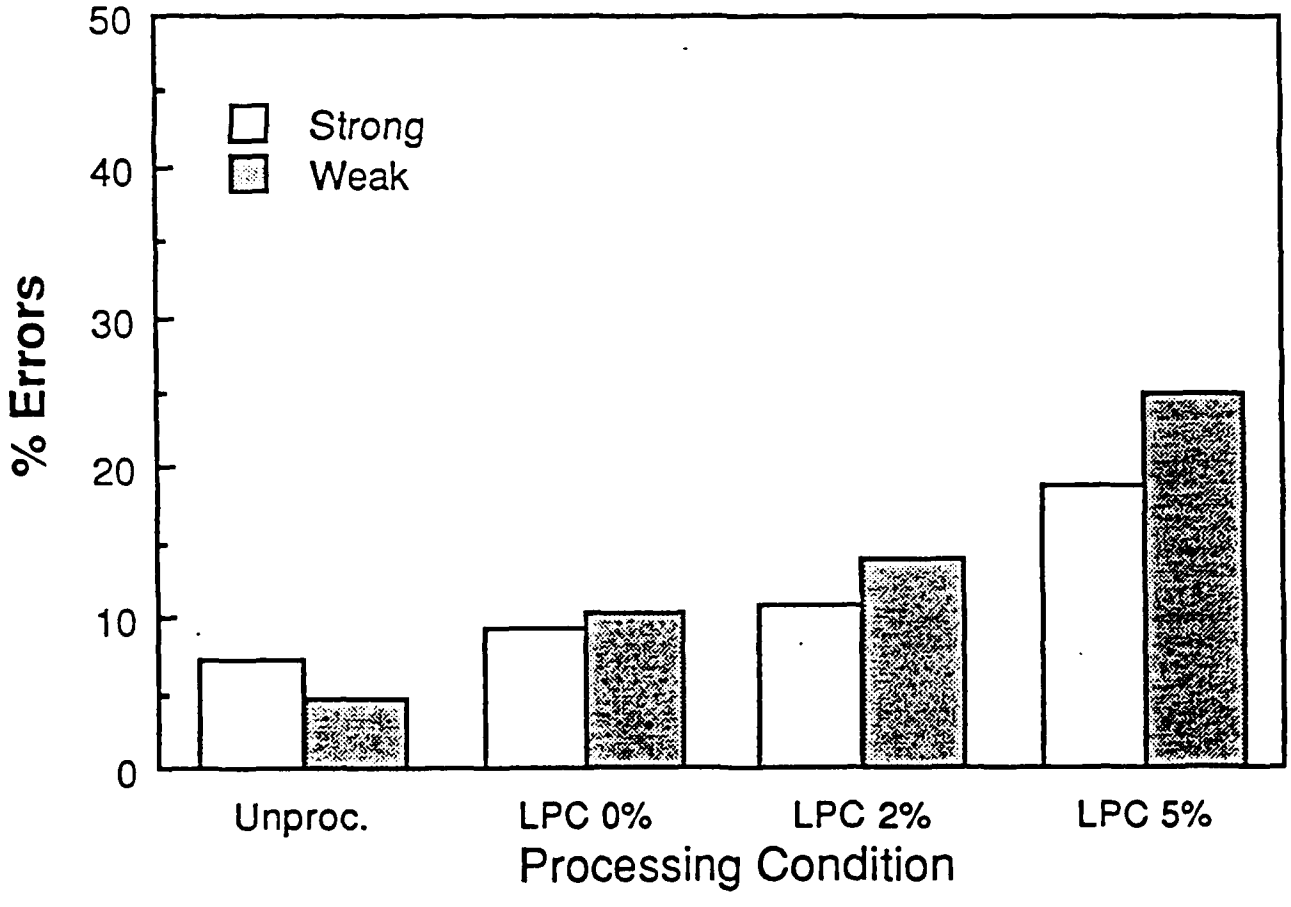

50
40
30
20
10
0
% Errors
Strong
Weak
Unproc.
LPC 0%
LPC 2%
LPC 5%
Processing Condition

Fig 1

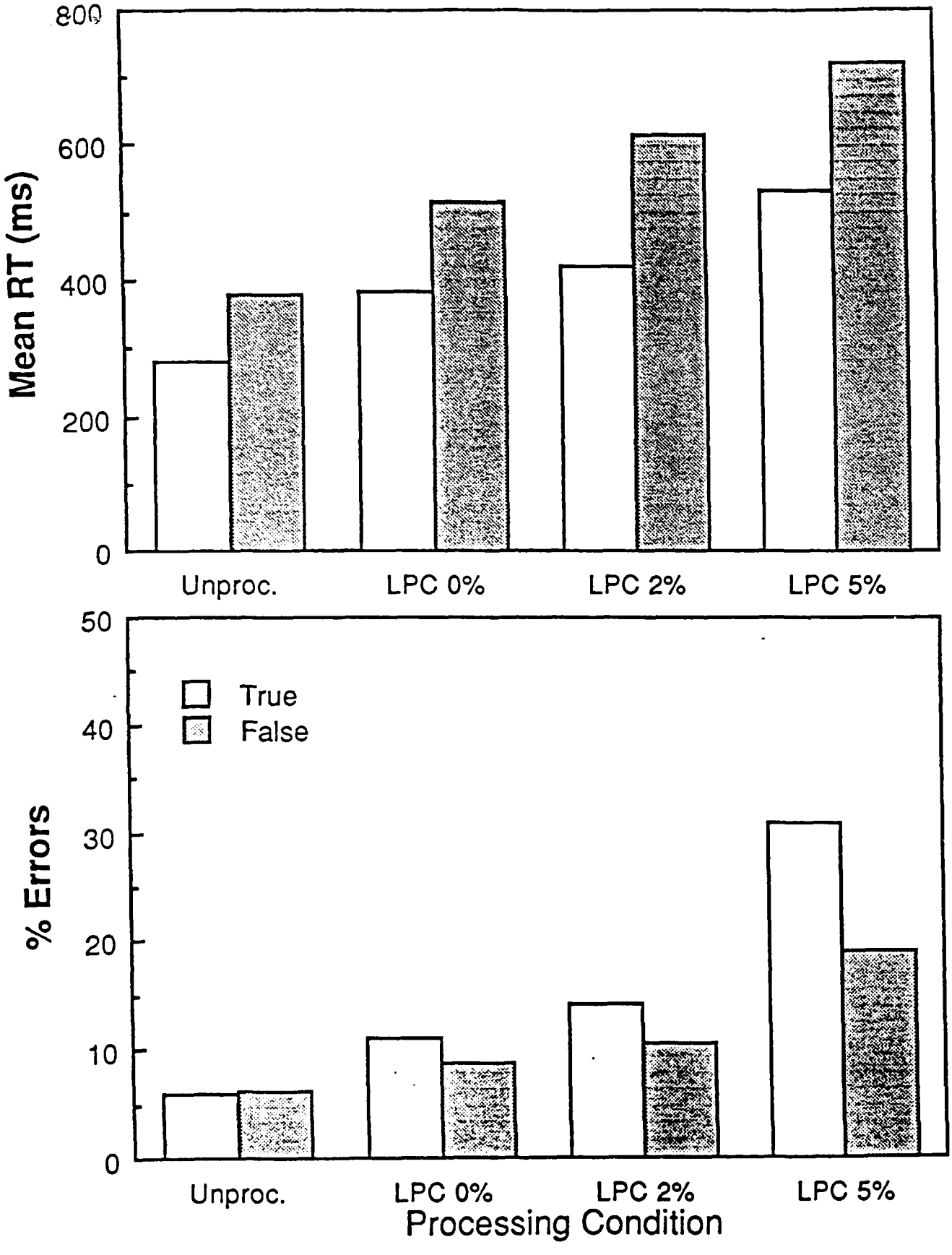

800
600
400
200
0
Mean RT (ms)
Unproc.
LPC 0%
LPC 2%
LPC 5%
50
40
30
20
10
0
% Errors
True
False
Unproc.
LPC 0%
LPC 2%
LPC 5%
Processing Condition

Fig 2

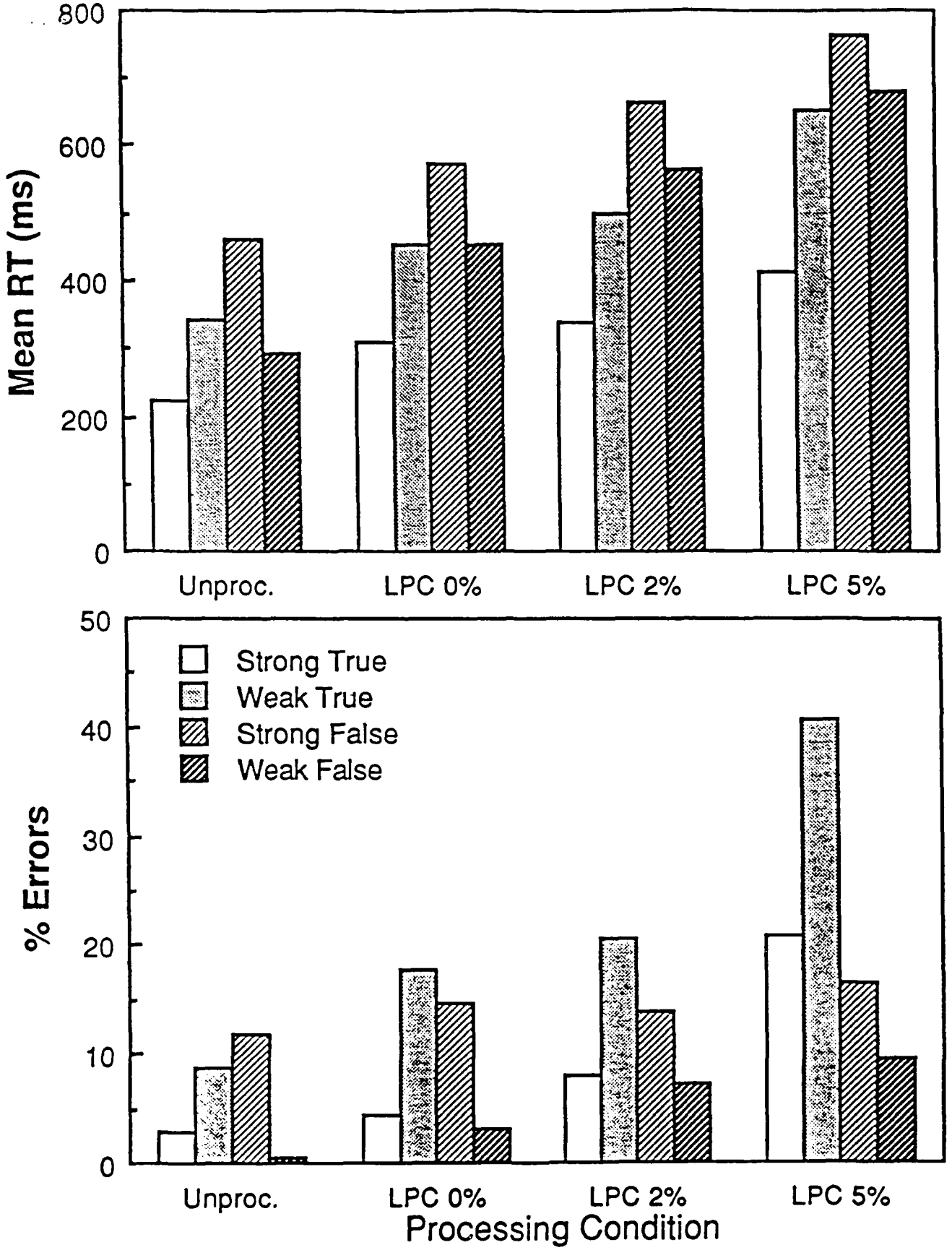

800
600
400
200
0
Mean RT (ms)
Unproc.
LPC 0%
LPC 2%
LPC 5%
50
40
30
20
10
0
% Errors
Strong True
Weak True
Strong False
Weak False
Unproc.
LPC 0%
LPC 2%
LPC 5%
Processing Condition

Fig 3

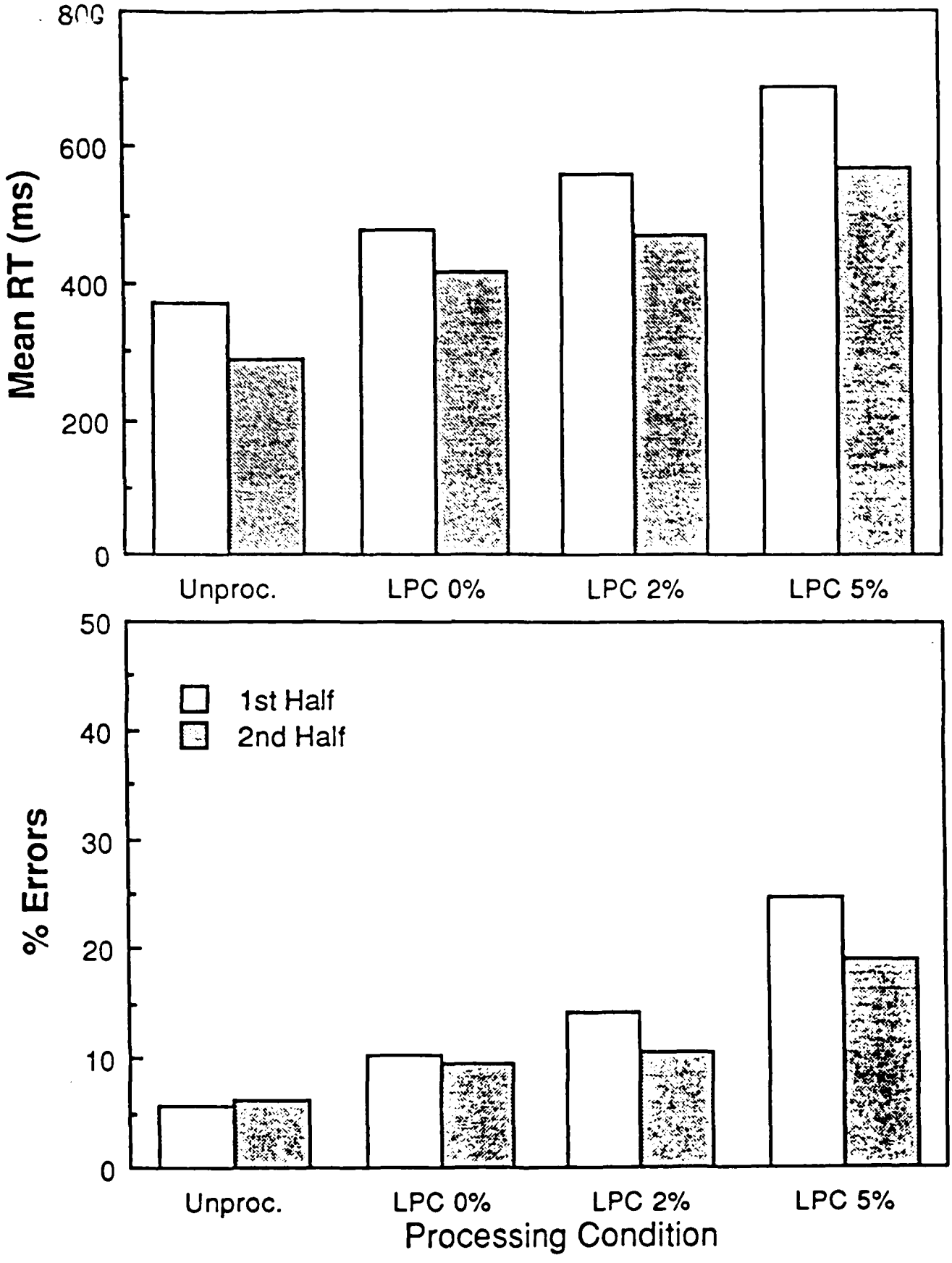

800
600
400
200
0
Mean RT (ms)
Unproc.
LPC 0%
LPC 2%
LPC 5%
50
40
30
20
10
0
% Errors
1st Half
2nd Half
Unproc.
LPC 0%
LPC 2%
LPC 5%
Processing Condition

Fig 4

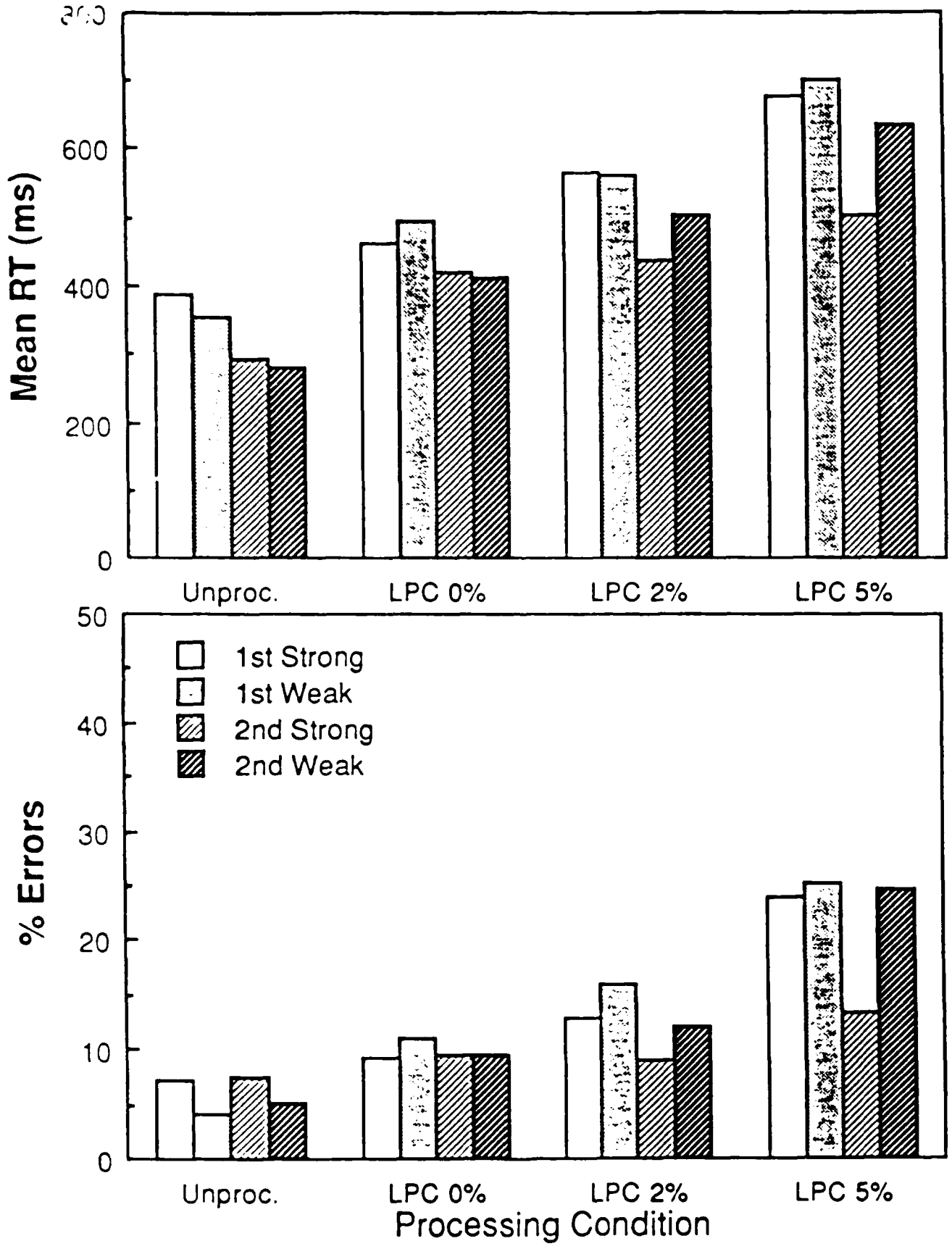

Mean RT (ms)
800
600
400
200
0
Unproc.
LPC 0%
LPC 2%
LPC 5%
50
40
30
20
10
0
% Errors
1st Strong
1st Weak
2nd Strong
2nd Weak
Unproc.
LPC 0%
LPC 2%
LPC 5%
Processing Condition

Fig 5

END

12-87

DTIC