RESPONSE TIMES TO SENTENCE VERIFICATION TASKS (SVTs) AS A MEASURE OF EFFORT IN SPEECH PERCEPTION

C. J. James, M.F. Cheesman, L. Cornelisse and L.T. Miller.

Hearing Health Care Research Unit
Department of Communicative Disorders
University of Western Ontario

ABSTRACT - Three studies are reported that assessed the potential use of three speech tests for obtaining a measure of listener effort. The tests were designed to yield two performance measures, response time and percent correct responses. The sensitivity of each test and each measure to changes in listener performance (practice effects) and listening condition (signal-to-noise ratio) was compared.

INTRODUCTION

Sentence verification tasks (SVTs) have been used by Pisoni, Manous and Dedina (1986) to subjectively measure speech synthesiser quality and by Gatehouse and Gordon (1990) to measure the benefit from amplification with hearing aid users. In these tasks, listeners are presented a sentence and then must report whether the sentence was true or false or repeat the stimulus verbatim. Response accuracy, scored as percent correct, and response times (RTs) are measured for each response. Both Pisoni et al. (1986) and Gatehouse and Gordon (1990) reported that response times were more sensitive to changes in listening conditions than percent correct scores.

The use of response times as an alternative, or supplement, to percent correct performance measures can overcome an inherent limit on the range of test sensitivity of percent correct scores in the measurement of speech intelligibility performance. For any test in which test items can be scored as correct and incorrect, the range of sensitivity is bounded by chance performance, at the lower end of the range, and by perfect performance at the top of the range. This range of performance extends over a very limited set of listening conditions, as illustrated in the psychometric curve in Figure 1. Figure 1 is a plot of the hypothetical intelligibility (percent correct score) as a function of the listening condition, in this case, the RMS level of the speech test materials. The slope of the psychometric function varies among speech tests; nonsense words yield a relatively shallow curve, having a slope of approximately 3% per decibel (Cheesman, Lawrence and Appleyard, 1992) and continuous discourse produces a very steep curve, with a slope of approximately 14% per decibel (Speaks, Parker, Harris and Kuhl, 1972). Percent correct scores can, therefore, be used to compare performance only over a relatively narrow range of listening conditions and, in this sense, are insensitive to changes in listening condition beyond this range.

Response time measures are less restricted than percent correct scores in terms of their range of measurement sensitivity. RTs generally decrease as the task becomes easier (and percent correct performance increases), as shown in Figure 1. For an RT measure to add to the information provided by percent correct scores, it must continue to decrease beyond the point at which percent correct performance asymptotes. The aim of the current set of experiments was to generate a set of stimuli that was sufficiently sensitive and reliable to measure an individual benefit benefit from small changes in listening condition and exhibited a wider range of sensitivity than traditional percent correct scores.

We developed three sets of stimuli: a set of conventional true/false type statements, synonym/non-synonym pairs and simple questions. The latter two types have not been previously considered for use in SVT type tasks, but have been used in assessments of basic cognitive ability. Two fundamental requirements of the stimuli were that the whole stimulus must be heard and that some linguistic or higher-level processing is required (i.e., the task could not rely on simple repetition or detection) in order for the listener to perform above chance. The

test items were designed such that the final word in the item must be heard in order to respond correctly to the item. It was hoped that this would prevent listeners from responding before the item had been completed. For these stimuli, statements must be judged true or false (categorically), synonym/non-synonym pairs must be judged synonymous or not and questions must be answered correctly (see examples provided below). One additional criterion used in the stimulus design was that the tests must be able to be administered repetitively with the same listener to obtain a measure of speech listening effort where intelligibility is very high.

To this end it was necessary to determine the difficulty of individual stimuli such that our final set of test items contained only stimuli that are easy to hear in good listening conditions and which listeners can respond to with a high degree of accuracy.

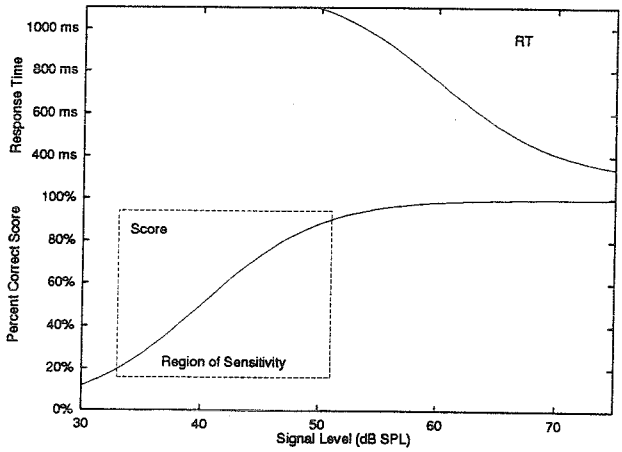

Figure 1: Hypothetical performance on a sentence verification task. Percent correct score and median response time (RT) versus presentation level are plotted. The limited range of sensitivity of the percent correct score is indicated; the RTs change over a wider range of testing conditions.

STIMULI

Stimulus set 1: Questions

The first set of test materials consisted of 72 simple questions with "yes", "no" or numerical answers. These were chosen to be unambiguous, easy questions, but required attention and audition of more than one word in each question. It should be noted that some of these questions pertain to subject specific information and therefore the actual accuracy of the response it now really known. Examples are:

- DO YOU HAVE A CAR / CAT / COAT / BIKE ? (20 items)
- HOW MANY LEGS HAS A CAT / HORSE / ANT / SPIDER ? (16 items)
- WHAT COMES AFTER / BEFORE TWO / THREE / FOUR ? (16 items)
- DO YOU LIKE CHEESE / HORSES / CARS / JAZZ ? (12 items)
- CAN YOU SPEAK FRENCH / DUTCH / ENGLISH / GERMAN ? (4 items)
- ARE SUMMERS / WINTERS HOT / COLD ? (4 items)

Stimulus set 2: Sentence Verification

A set of 332 true or false statements was constructed; these are similar to the type of sentence verification items used by Pisoni et al (1986). Subjects were asked to respond with "true" or "false". The sentences were expected to vary in their intrinsic difficulty, because of the differences in their length, but were designed to be easily interpreted. The test items were constructed from a limited vocabulary. Many vocabulary items served as adverbs, adjectives and nouns, depending on their position in the sentences, this prevented listeners from responding to a single word in the subject phrase; rather, the listeners frequently needed to attend to the complete subject phrase in order to identify the subject in the sentence. The large number of test items was expected to minimize potential learning effects. Examples are:

- A HORSE IS AN ANIMAL.
- A HORSE BARN IS A BUILDING.
- AN ORANGE CAR IS A VEHICLE.
- CATS ARE VEGETABLES.
- A HORSE BARN IS AN ANIMAL.
- ORANGE HOUSE COATS ARE PETS.

Stimulus set 3: Synonyms

A set of 297 synonym and non-synonym pairs were constructed. The synonym pairs were selected from a set of published norms for synonyms (Wilding and Mohindra, 1983a, 1983b). Only those norms reported as the most frequently used and acceptable were chosen as items. Non-synonym pairs were randomly selected from the set of synonym pairs and checked for disambiguity. Listeners could answer "true" or "false", or "yes" or "no" to each pair. Examples are:

- SCENT - SMELL
- SORROW - SADNESS
- RUG - MAT
- COST - MAT
- LABOUR - CONTENT
- DANGER - COPY

Recording

The stimuli were produced by a female Canadian English talker and digitally recorded. Throughout the recording session, the talker's vocal effort was monitored on a VU meter in order to maintain a consistent overall level. No further adjustments were made to the stimulus levels. The words in each synonym or non-synonym pair were recorded separately. A delay of 500 ms was inserted between the first and second word of each word pair in the Synonym test items. Sentences and Questions were recorded in their entirety.

METHODS

Three investigations of the properties of these stimuli have been undertaken. These were designed to obtain data relating the difficulty, consistency and sensitivity of stimuli in terms of correct responses and RT. One experiment measured learning effects overall a period of several days, the second examined the sensitivity and long term reliability, and the third measured short-term practice effects.

Response Mode

Pisoni, Manous and Dedina (1986) both obtained RTs from manual responses, and Gatehouse and Gordon (1990) used vocal responses. The particular test materials used by Gatehouse and Gordon precluded the use of a manual response; however, Baer, Moore and Gatehouse (1993) modified the test materials so that they were condusive to manual responses. Because the subjects in these earlier studies were given only two response alternatives, a manual response mode, using two response buttons avoided the problems of poor accuracy and false triggering associated with voice-activated switches. A vocal response time system was devised for use with our test materials that was designed to be less influenced by artifact than these earlier systems. This system is described in detail by James (1994). With this system, responses are digitally recorded and the resulting audio files are synchronised with the presentation of stimuli. Response times are extracted from the response files using a highly robust "silence" detection algorithm. Response times reported here are relative to the end of the stimuli.

There has been some comparison of vocal to manual response times in monitoring experiments (Till, Reich, Dickey and Seiber, 1974). Here a button is pushed or a simple production (such as "uh") is made when a tone was turned off. The results suggest that button pushing is consistently faster (around 30 ms) compared to phonation, but that RTs are of similar variability. Thus there are no obvious disadvantages to using vocal responses. In the current study the use of vocal responses also enabled the use of an open response set (responses other than "yes" and "no" or "true" or "false" could be accepted).

The designs of three experiments are briefly described below. Experiments 1 and 2 looked at all three stimulus sets. Experiment 3 used only the Questions All stimuli were presented in sound-field in an audiometric booth to maintain parity with use for hearing aid evaluation. The subjects were asked to respond vocally, and as quickly and as accurately as they could.

Experiment 1: Learning Effects

Three subjects have been tested with all three sets of stimuli on five occasions with two to three days between testing sessions. Within each session, the three sets were administered in the same order, with approximately ten minutes between each set. On each occasion the stimuli were presented in quiet at around 50 dB SPL and in a different random order.

Experiment 2: Sensitivity

Five subjects participated a sensitivity study. On the first occasion all three sets of stimuli stimuli were presented at a comfortable level in quiet, i.e. an RMS speech level of 60 dB SPL. In subsequent test sessions noise was added at four different levels, resulting in signal to noise ratios of 15 dB, 10 dB, 5 dB and 0 dB RMS. These levels were chosen to obtain a range of 75%-100% percent correct scores. The order of testing the noise conditions was randomised for each listener.

Experiment 3: Practical Simulation

In addition to Experiments 1 and 2 some informal practical trials have been undertaken with the Questions. Four subjects have been tested to date. In these trials the stimuli were presented in each of the five noise conditions described above in order of decreasing signal to noise level in a single test session of approximately 25 minutes duration. This experiment examines the combined effects of short term learning, noise condition, and repetition of the task (fatigue).

RESULTS AND DISCUSSION

The focus of this study was to ascertain whether response times were more sensitive than accuracy scores for the conditions imposed. We hoped to see differences in response time could be detected under conditions where only marginal differences in percent correct scores are observed. The data from the first experiment for all stimulus sets showed trends of improvements in performance with practice, with some reduction in the variability of the RTs across successive sessions. The most marked reductions in RT were seen between the first and second sessions.

In Experiment 2, with both the Questions and Sentences, test correct scores were around 95-100% in all test conditions. The Synonyms showed a wider range of difficulty with performance ranging from around

75% at the lowest S/N to nearly 100the highest S/N. For the Synonyms a similar pattern was not observed with the RT measures. RT results for the Questions are shown in Figure 2. Subject 1 exhibits a pattern of increasing RT with decreasing signal to noise ratio. For the other subjects, although there is some increase in RT with reduced signal to noise ratio there is considerably greater variability. The data from Experiment 1, to some extent, can be helpful in understanding the source of the apparent difficulty of the quiet "Q" condition which was always used in the first session in Experiment 1. For both Experiments 1 and 2 the results from the Sentences and more variable and trends less apparent than those for the Questions and Synonyms. The fact that differences in response times are seen in Experiment 2 without similar differences in correct scores indicates that the Questions may be the most sensitive set.

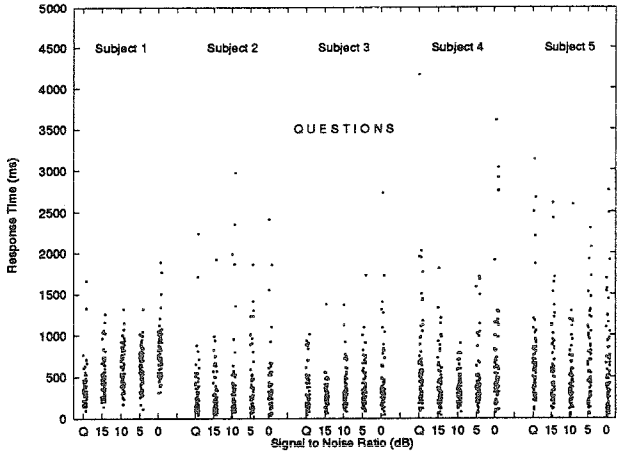

Figure 2: "Scatter-grams" showing the distribution of response times for each subject under 5 conditions with various signal to noise ratios. Subject 1 shows the clearest trend in increasing response time with decreasing signal to noise ratio.

In Experiment 3, the there were some clear effects of learning and conditions. Figure 3 shows results for two subjects tested on two occasions. On Day 2, a clearer effect of S/N on RT is apparent. We may assume that results for Day 2 are valid in that mean response times are faster than for Day 1, this illustrates a practice effect. The implications for practical use are that it may be necessary to include practice trials in order to stabilise performance or to quantify the learning effects and adjust for such effects in the data sets.

FUTURE CONSIDERATIONS

The primary aim here was to investigate the sensitivity possible from the SVT paradigm. That is we wish to demonstrate a consistent difference in response times (RTs) when the same stimuli are presented in different conditions. The data described here are undergoing further analysis in terms of the utility and consistency of individual stimuli. It is not clear that the increase in response times with reduction of signal to noise ratio will be of use in clinical testing. We have not, at this stage, compared sensitivity for these conditions with existing list-score tests such as mono-syllables and VCV tokens. The accuracy scores from the Synonyms may reflect those obtainable with, for example, spondaic words, or they may be a better set based on accuracy scores! We will also be investigating the sensitivity of the stimuli in terms of performance in various filtering and noise conditions which may be envisaged to be closer to differences in amplification characteristics provided by different hearing aids.

ACKNOWLEDGEMENTS

We would like to thank Dr. D.G. Jamieson for his advice. We are also grateful to Diahann Whitefield and Vicky Curtis for assistance with data collection and to the Ontario Ministry of Health and the Natural Sciences and Engineering Research Council of Canada for financial support.

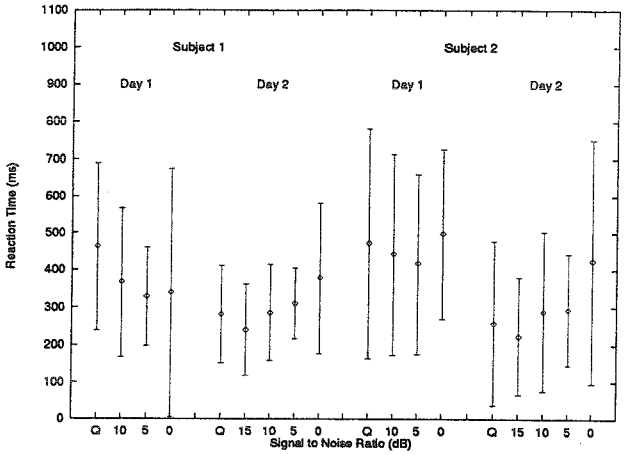

Figure 3: The performance of two subjects on two occasions. The data show a practice effect on Day 1 which is offset by condition effects. On Day 2, reaction times are generally faster and there is a clear effect of condition after the first first practice trial.

REFERENCES

Cheesman, M.F., Lawrence, S. and Appleyard, A. (1992). "Prediction of performance on a nonsense syllable test using the Articulation Index." In Ohala, J. J., Nearey, T. M., Derwing, B. L., Hodge, M. M., and Wiebe, G. E. (Eds.), ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing. Edmonton: University of Alberta. pp. 1123-1126.

Gatehouse, S. and Gordon, J. (1990) "Response times to speech stimuli as measures of benefit from amplification", British Journal of Audiology, 24, 63-68.

Baer, T., Moore, B.C.J, and Gatehouse, S. (1994) "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality and response times." Journal of Rehabilitation Research and Development, 30:1, 49-72.

James, C.J. (1994) "A verbal response time system for use with sentence verification tasks (SVTs)", Under Review.

Pisoni, D.B., Manous, L.M. and Dedina, M.J. (1987) "Comprehension of Natural and Synthetic Speech: The effects of predictability on the verification of sentences controlled for intelligibility.", Computer Speech and Language, 2, 303-320.

Speaks, C., Parker, B., Harris, C. and Kuhl, P. (1972) "Intelligibility of connected discourse", Journal of Speech and Hearing Research, 15, 590-602.

Till, J.A., Reich, A., Dickey, S. and Seiber, J. (1983) "Phonatory and manual reaction times of stuttering and nonstuttering children.", Journal of Speech and Hearing Research, 26, 171-180.

Wilding, J. and Mohindra, N. (1983a) "Ratings of the degree of synonymity of 279 noun pairs.", British Journal of Psychology, 72, 231-239.

Wilding, J. and Mohindra, N. (1983b) "Preferred synonyms for each of 279 noun pairs.", British Journal of Psychology, 74, 91-106.