

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 25 (2001-2002)
Indiana University

**Effects of Response Format in Spoken Word Recognition Tests:
Speech Intelligibility Scores Obtained from Open-Set,
Closed-Set, and Delayed Response Tasks¹**

Cynthia G. Clopper and David B. Pisoni

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by NIH-NIDCD Research Grant DC00111 and Training Grant DC00012 to Indiana University.

Effects of Response Format in Spoken Word Recognition Tests: Speech Intelligibility Scores Obtained from Open-Set, Closed-Set, and Delayed Response Tasks

Abstract. In studies of spoken word recognition, factors such as the materials to be recognized and the degradation of the speech signal have been discussed at length in the literature. The factor that has been largely ignored, however, is that of response format. In particular, word recognition studies in the last forty years have been largely closed-set tasks, despite the fact that the original tests used by military personnel for testing radio equipment during World War II were designed to be open-set. Recent findings suggest that open-set and closed-set tests may differ in more than just their level of chance performance. Sommers, Kirk and Pisoni (1997) found that lexical competition and talker variability do not produce reliable effects on word recognition scores in closed-set tasks, but they do produce robust effects on performance in open-set tasks. The present study was designed to replicate the earlier findings reported by Sommers et al. and explore the response demands in a modified closed-set task. Specifically, one group of listeners participated in a closed-set word recognition task in which the response alternatives were not provided until one second after the offset of the signal. Their performance was compared to listeners in an open-set task and a traditional closed-set task. Results revealed effects of lexical competition only for the listeners in the open-set task. We assume that in open-set tests, the listener must access the lexicon. These findings therefore suggest that even a delay of one second is not adequate to force the listener to process the stimuli in a way that is similar to open-set processing. The major implication of these results is that even modified closed-set tasks cannot replace open-set tasks as valid measures of word recognition performance.

Introduction

Word recognition tasks have been used for more than a half-century in settings as diverse as the testing of military radio equipment (Miller, 1946a), studies of speech intelligibility (Horii, House, & Hughes, 1970), clinical tests of the auditory capabilities of hearing-impaired individuals (Owens, Kessler, Telleen, & Schubert, 1981), and studies of the cognitive implications of word superiority effects in letter recognition tasks (Reicher, 1969; Wheeler, 1970).

In the word recognition literature, factors such as mode of recognition (visual or auditory), the materials to be recognized, the content of the foils presented as response alternatives, and the type of masking or degradation have been manipulated by researchers to better suit the particular goals of their experiments. In spoken word recognition, Miller (1946a) preferred to use nonsense syllables because he was most interested in the accurate transmission of segmental information in speech. However, he also found that when using a limited set of real English words, speech intelligibility performance increased if the words were multisyllabic and phonetically distinct (Miller, 1946b).

House, Williams, Hecker, and Kryter (1965) used sets of phonetically similar real words because they were interested in how well listeners could differentiate words in English based on single consonant differences. Foster and Haggard (1987) also used real words but selected response alternatives based on binary feature distinctions instead of mere consonantal contrasts. More recently, Pisoni and his colleagues (Kirk, Pisoni, & Osberger, 1995; Luce & Pisoni, 1998; Meyer & Pisoni, 1999; Mullennix, Pisoni, &

Martin, 1989) have shown that the lexical properties of the target items (e.g., frequency, familiarity, and neighborhood density) as well as talker variability can affect spoken word recognition performance in both normal-hearing and hearing-impaired listeners.

Miller (1946a, b) could rely on the noise introduced through the radio equipment to degrade his stimuli, but noisy conditions have been simulated for normal-hearing listeners in other studies using a wide variety of methods including white noise (Sommers et al., 1997), envelope-shaped white noise (Horii et al., 1970), simulated airplane noise (Black, 1957), multi-talker babble (Nakanishi, 1988), and bit-flipping (Saldana, Pisoni, Fellowes, & Remez, 1996).

While these signal manipulations, as well as the clinical differences between normal-hearing and hearing-impaired populations, have been studied and discussed extensively in the literature, the issue of the nature of the task itself has received relatively little attention. Most of the early speech intelligibility tests described by Miller (1946a) used open-set tests of word (or syllable) recognition. However, other researchers began to switch to multiple choice speech intelligibility tests for speed and efficiency (Black, 1957). The underlying assumption of the new closed-set tests was that the process of word recognition would be the same, regardless of the response format of the test. It was thought that the only difference between open-set and closed-set tasks was chance performance ($1/\infty$ in open-set tests; $1/N$ in closed-set tests, where N is the number of response alternatives).

In closed-set tests, researchers have varied the response alternatives in consideration of phonological properties to examine more closely the effects of linguistic properties such as phoneme confusability on word recognition. In designing one of the earliest closed-set speech intelligibility tests, Black (1957) selected three foils for each target word from incorrect responses to the targets presented in an open-set condition in noise. That is, he used foils that he knew to be confusable with his targets based on data collected in a similar task with human participants. For example, a target word *burst* might be accompanied by the foils *nurse*, *first*, and *birth* on the response sheet.

House et al. (1965) used a different approach to create their six-alternative forced choice speech intelligibility test. They created 50 sets of six CVC words each. In 25 of the sets, all of the words in each set contained the same initial consonant and the same vowel. For example, *bat*, *bad*, *back*, *bass*, *ban*, and *bath* are one set in which only the final consonant varies between the words. In the remaining 25 sets, all of the words contained the same vowel and the same final consonant. An example of this kind of set might include *led*, *shed*, *red*, *bed*, *fed*, and *wed* in which all of the words rhyme but have different initial consonants. All of the foils then differed from the target word by only a single consonant (either the initial or the final). The result of this kind of design is that any word in a given set could be used as the target word and the other words in the set would be its foils.

In the Minimal Auditory Capabilities (MAC) battery, Owens et al. (1981) used a similar design for determining phoneme discrimination abilities. In a series of four-alternative forced choice tests, Owens et al. investigated word recognition abilities for English CVC words differing either in initial consonant (e.g., *din*, *bin*, *fin*, and *gin*), final consonant (e.g., *rid*, *rip*, *rib*, and *ridge*), or vowel (e.g., *fool*, *full*, *fall*, and *foul*).

Finally, Foster and Haggard (1987) considered an even smaller unit of linguistic contrast in their creation of the Four Alternative Auditory Feature (FAAF) test. Instead of varying their foils and targets based on phonemic consonantal contrasts, Foster and Haggard built lists of targets and foils that differed only on individual featural dimensions based on minimal pairs. For example, the target word *mail* (with an initial bilabial nasal) might be accompanied by the foils *bail* (with an initial bilabial stop), *nail* (with an initial alveolar nasal), and *dale* (with an initial alveolar stop).

Despite the many systematic considerations of the content of the targets and the foils, the effects of response format on spoken word recognition performance have not been considered. Pollack, Rubenstein, and Decker (1959) reported that differences in word frequency effects were observed between known (closed-set) and unknown (open-set) word sets in speech intelligibility tests. In particular, word frequency affected performance on the unknown sets of words but not the known sets of words. This effect was found regardless of the number of words in the set, which ranged in size from 8 to 144 words. These early results suggest that some aspects of the normal word recognition process, such as those responsible for word frequency effects in recognition, may be bypassed in closed-set speech intelligibility tests.

In a recent study from our laboratory, Sommers et al. (1997) used both an open-set and a closed-set task in evaluating the word recognition performance of cochlear implant users. They manipulated lexical competition (based on neighborhood density and neighborhood frequency), talker variability, and response format in both normal-hearing listeners (in both quiet and noisy conditions) and cochlear implant users (in a quiet condition only). Sommers et al. found that performance on the closed-set task was better than performance on the open-set task for both groups of listeners. However, the effects of lexical competition and talker variability were observed only in the open-set condition. These effects were found in both groups of listeners. No significant effects were found for the normal-hearing listeners in the quiet condition because their performance was at ceiling for all variables.

The findings by Sommers et al. (1997) also suggest that the processes used in recognizing spoken words in closed-set test formats may not be equivalent to the processes used in open-set word recognition. In particular, it is assumed in most models of word recognition that the lexicon is accessed through some kind of activation process based on the properties of the acoustic signal (Jusczyk & Luce, 2002). While open-set tasks using isolated single words may not be a perfect model of real-life language situations that typically involve the use of sentences, the contribution of semantic context, and lower levels of degradation, closed-set word recognition tasks may fundamentally bypass some of the higher-level lexical access processes that are certainly at work in everyday language situations. Under these listening conditions factors such as lexical competition and talker variability may come into play.

Given that closed-set tests are easier to administer and score, it would be useful in clinical settings to have a closed-set test to measure word recognition performance. However, given that open-set tests may employ a fundamentally different set of cognitive processes, scores on closed-set tests in clinical situations might provide an inflated measure of the listener's speech perception skills. The goal of the present study was to determine if a modified closed-set task could produce the same effects of lexical competition and talker variability as an open-set test. In particular, the closed-set task used in this study was modified so that the listeners did not see the response alternatives until 1000 ms after the presentation of the spoken test word. We predicted that this delay in the response alternatives might result in the activation of the normal (i.e., open-set) word recognition process, thus revealing the same effects found in open-set tests. One group of young normal-hearing listeners participated in the delayed closed-set condition. Two additional groups of listeners were used for comparison: one was assigned to an open-set condition and the other was assigned to a more traditional closed-set condition in which the response alternatives were provided before the presentation of the stimulus word.

Experiment 1: Level of Signal Degradation

In order to determine appropriate levels of signal degradation, a pilot study was carried out using three levels of signal degradation in two conditions: open-set and closed-set. The goal of this preliminary

study was simply to determine which levels of signal degradation would allow for performance above floor in the open-set condition and below ceiling in the closed-set condition.

Methods

Materials. A set of 132 CVC English words were selected for use in this study from the Modified Rhyme Test (MRT; House et al., 1965) and the Phonetically Balanced (PB; IEEE, 1969) word lists. The words were divided into two groups based on measures of lexical competition ('easy' and 'hard') with 66 words in each group. Based on the word counts in Kučera and Francis (1967), mean log frequency was equated across the two groups of words. Similarly, mean word familiarity, as judged by Indiana University undergraduates (Nusbaum, Pisoni, & Davis, 1984), was also equated across the two groups. The lexically 'hard' words had a significantly higher mean lexical density (defined as all words that differ from the target by a single phoneme substitution, deletion, or addition) and the neighbors of the 'hard' words had a significantly higher mean log frequency than the neighbors of the lexically 'easy' words. As in Sommers et al. (1997), the 'easy' and 'hard' sets of test words were defined based on these differences in mean density and mean neighborhood log frequency. Table 1 shows a summary of the means for each set of words on the four lexical measures, as well as significance values for the t-tests run to compare the means.

	Easy	Hard	
Mean Log Frequency	1.91	2.02	$p = .41$
Mean Familiarity	6.81	6.70	$p = .27$
Mean Density	16.03	24.86	$p < .001$
Mean Neighborhood Log Frequency	1.83	2.17	$p < .001$

Table 1. Lexical properties of 'easy' and 'hard' words.

Five male talkers were selected from a total of 20 talkers (10 males and 10 females) who were recorded reading the MRT and PB word lists for the PB/MRT Word Multi-Talker Speech Database (Speech Research Laboratory, Indiana University). Based on the results of a similarity judgment task (Goh, 2001), a clustering analysis was computed for the male and female talkers. In order to ensure that the selected talkers could be discriminated from one another, two talkers were taken from two of the major clusters and the fifth was taken from the third cluster in the clustering solution. In the two-dimensional multi-dimensional scaling solution to the similarity judgments, the five selected talkers were well-dispersed in both dimensions of the space.

Each of the 132 words was spoken by each of the five talkers, for a total of 660 tokens. Each token was stored in an individual sound file in .wav format. For the present study, the tokens were degraded using a bit-flipping procedure written in Mathworks Matlab. In bit-flipping, degradation is introduced into the signal by flipping the sign of a randomly selected proportion of the bits in the signal. The higher the percentage of bits that are flipped, the more degraded the signal is. In this case, three levels of degradation were selected: 10%, 20%, and 30%.

For the closed-set condition, a six-alternative forced-choice task was designed. Each of the five foils for each target word differed from the target by the substitution of a single phoneme. Ideally, all of the foils were rated by undergraduates as having a familiarity rating of greater than 6.0 on a 7-point scale (Nusbaum et al., 1984). In addition, the goal was to have two foils higher in frequency, two foils lower in frequency, and one foil the same frequency as the target. Finally, two foils differed from the target with

respect to the initial consonant, two with respect to the final consonant, and one with respect to the vowel. This general schema for selecting foils could not be followed in all cases due to the constraints of the English language. Therefore, some sets of foils did not match the schema with respect to minimum familiarity, frequency distributions, or phoneme substitution location, but in all cases the foils differed from the targets by the substitution of a single phoneme and the familiarity of all foils was greater than 5.0 based on scores obtained from the Hoosier Mental Lexicon database (Nusbaum et al., 1984). A complete list of targets and foils is shown in Appendix A.

Listeners. Twenty-one Indiana University undergraduates, 10 males and 11 females, participated as listeners in this experiment. Ten listeners were assigned to the open-set condition and 11 listeners were assigned to the closed-set condition. All listeners were native speakers of English, with no history of hearing or speech disorders reported at the time of testing. The listeners received partial course credit for participating in the experiment.

Procedure. The listeners were seated at personal computers equipped with Beyerdynamic DT100 headphones. All 132 words were presented one time at each of the three levels of degradation, for a total of 396 trials per listener. All of the words at one level of degradation were presented in random order within a single block. Presentation order of the degradation blocks was balanced across listeners. Each listener heard only one of the five talkers and that talker remained constant across all three blocks. The talkers were randomly assigned to the listeners such that there were two or three listeners per talker per listener group.

The listeners assigned to the open-set condition heard the words presented one at a time over the headphones at 75dB and were simply asked to type in the word that they thought they had heard using a standard keyboard. The next trial was initiated by pressing a “Next Trial” button on the computer screen by clicking on it with the mouse.

The listeners assigned to the closed-set condition also heard the words presented one at a time over the headphones at 75dB. Simultaneously with the onset of the auditory presentation of each word, six response alternatives were presented on the computer screen. After hearing the word, listeners were asked to use the mouse to select which one of the six response alternatives they thought they had heard. The next trial was initiated by pressing a “Next Trial” button on the computer screen by clicking on it with the mouse.

Prior to data analysis, open-set responses were corrected by hand for obvious typographical errors in cases where the response given was not a real word and for homonyms such as *pear* for *pair*.

Results

Figure 1 shows the mean performance by the two groups of listeners for each of the three levels of degradation. A two-way ANOVA (degradation x response format) confirmed the significant effect of degradation ($F(2, 60) = 68.4, p < .001$). Post-hoc Tukey tests revealed that performance at 10% degradation was significantly better than performance at 20% and 30% degradation ($p < .001$ for both) and that performance at 20% degradation was significantly better than performance at 30% degradation ($p < .001$). In addition, there was a significant main effect of response format ($F(1, 60) = 350.3, p < .001$). Closed-set performance was always better than open-set performance, as expected based on the results of Sommers et al. (1997). The interaction between response format and level of degradation was not significant.

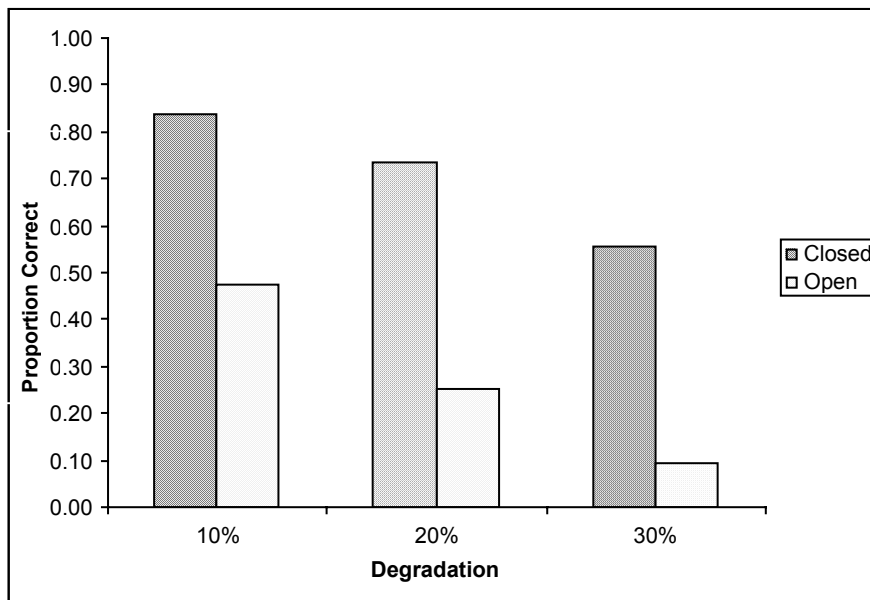


Figure 1. Word recognition performance as a function of response format and level of degradation.

Discussion

Given that open-set performance was close to floor (9%) at 30% degradation and that closed-set performance was not at ceiling (84%) at 10% degradation, 10% and 20% levels of degradation were selected for use in the next experiment which manipulated the closed-set response format.

Experiment 2: Response Format Task Analysis

The goal of this study was to determine if a modified closed-set test of speech intelligibility could be used as an alternative to open-set tests in clinical settings. Therefore, a delayed response closed-set task ('after' condition) was used as the test condition. In addition, open-set ('open' condition) and traditional closed-set ('before' condition) tasks were used as control conditions. Response condition was a between-subject variable. Other variables that have been shown to reveal differences between open-set and closed-set word recognition tests were manipulated within subjects: level of signal degradation, talker variability, and lexical competition.

Methods

Materials. The same test words, spoken by the same five male talkers in Experiment 1, were used in this experiment. In addition, the same response alternatives were used in the two closed-set conditions as in Experiment 1. In this experiment, only 10% and 20% levels of degradation were used.

Listeners. Eighty-one Indiana University undergraduates, 26 males and 53 females, served as listeners in this experiment. All received partial course credit for participating. Data from nine listeners were excluded from the final data analysis because the listeners were bilingual (four participants), were non-native speakers of English (one participant), had a history of speech or hearing disorder (three participants), or fell outside the age distribution of the other participants (one participant). Data from 24

listeners in each of the three conditions are reported in the final analyses below. All 72 of these listeners were monolingual native speakers of English with no reported history of hearing or speech disorders.

Procedure. The listeners were seated at personal computers equipped with Beyerdynamic DT100 headphones. Each of the 132 words were presented one time to each listener. The procedure was divided into a practice block and four test blocks. The practice block consisted of two randomly selected ‘easy’ words and two randomly selected ‘hard’ words spoken by the same talker presented at 10% degradation. Listeners completed the four practice trials and were encouraged to ask questions regarding the procedure before continuing on to the test blocks. Each of the four test blocks contained 16 ‘easy’ and 16 ‘hard’ words. The four blocks represented all possible combinations of level of degradation (10% or 20%) and number of talkers (single or multiple). For each listener, one talker was selected as the single talker and that talker was only used in the practice block and the two single talker blocks. The remaining four talkers were used in the multiple talker blocks. The words in each block were randomly and exhaustively selected for each listener, such that each listener heard each word only once, but the talker and block in which each word appeared was random. Single talkers and block orders were balanced across listeners to ensure that any observed effects were not due to characteristics of a specific talker.

As in Experiment 1, the listeners in the open-set condition heard the words presented one at a time over headphones at 75dB and were asked to type in the word that they thought they had heard using a standard keyboard. The next trial was initiated by pressing a “Next Trial” button on the computer screen by clicking on it with the mouse.

The listeners in the ‘before’ closed-set condition heard the words presented one at a time over headphones at 75dB. One second prior to the onset of the auditory presentation of the words, the six response alternatives were presented in random order in a single row on the screen. After hearing the word, listeners were asked to use the mouse to select which one of the six response alternatives they thought they had heard. The next trial was initiated by pressing a “Next Trial” button on the computer screen by clicking on it with the mouse.

The listeners in the ‘after’ closed-set condition also heard the words presented one at a time over headphones at 75dB. However, in this condition, the six response alternatives were presented on the screen one second after the end of the auditory presentation of the words. After the presentation of the response alternatives, listeners were asked to use the mouse to select which one of the six response alternatives they thought they had heard. The next trial was initiated by pressing a “Next Trial” button on the computer screen by clicking on it with the mouse.

Prior to the analysis of the data, open-set responses were corrected by hand for obvious typographical errors and homonyms, using the same criteria as in Experiment 1.

Results

Response Format. We observed an overall effect of response format on performance, collapsed across all of the other independent variables: level of degradation, talker variability, and lexical competition, as shown in Figure 2. A one-way ANOVA revealed a significant main effect of response format ($F(2, 285) = 279.6, p < .001$). Post-hoc Tukey tests revealed that performance in the open-set condition was significantly worse than performance in either of the closed-set conditions ($p < .001$ for both) and that performance in the ‘after’ condition was significantly worse than in the ‘before’ condition ($p < .001$).

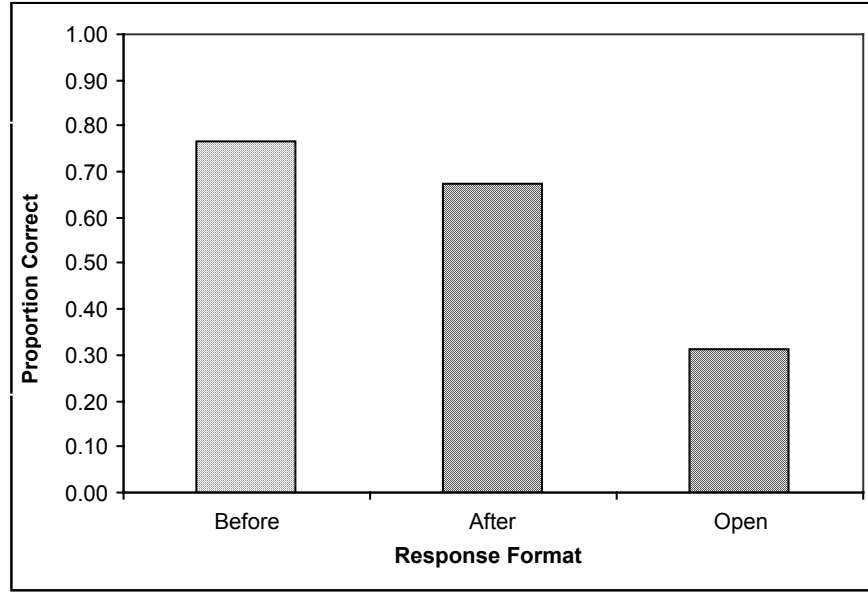


Figure 2. Word recognition performance as a function of response format, collapsed across level of degradation, talker variability, and lexical competition.

Signal Degradation. As shown in Figure 3, an effect of level of degradation was observed for all three response formats, collapsed across talker variability and lexical competition. A two-way ANOVA (response format x degradation) revealed a significant main effect of response format ($F(2, 141) = 479.0$, $p < .001$). Post-hoc Tukey tests again confirmed that performance in the ‘before’ condition was significantly better than performance in the ‘after’ and open-set conditions ($p < .001$ for both) and that performance in the ‘after’ condition was better than in the open-set condition ($p < .001$). There was also a significant main effect of degradation ($F(1, 141) = 235.3$, $p < .001$). Performance at low degradation (10%) was better than performance at high degradation (20%). Planned post-hoc t-tests revealed a significant effect of degradation for all three conditions: $t(23) = 9.5$, $p < .001$ for the ‘before’ condition; $t(23) = 10.4$, $p < .001$ for the ‘after’ condition; $t(23) = 15.5$, $p < .001$ for the open-set condition. The response format x signal degradation interaction was also significant ($F(2, 1) = 4.7$, $p = .01$). In order to determine the location of the interaction, a one-way ANOVA on the mean differences in performance between the two levels of degradation for the three conditions was calculated. The ANOVA revealed a significant effect of condition ($F(2, 69) = 8.2$, $p < .001$). Post-hoc Tukey tests revealed that the effect of degradation was significantly lower for the ‘before’ condition than either the open-set condition or the ‘after’ condition ($p < .001$ and $p < .05$, respectively). The effect was not significantly different between the ‘after’ condition and the open-set condition.

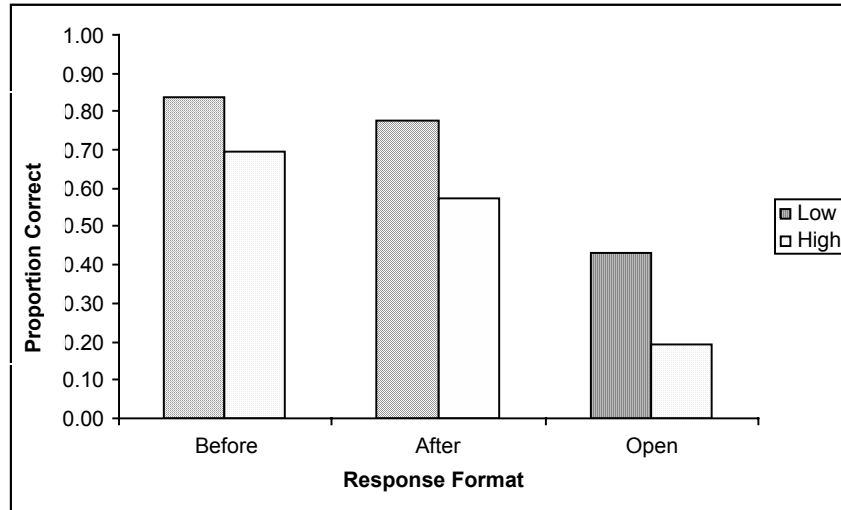


Figure 3. Word recognition performance as a function of response format and level of degradation, collapsed across talker variability and lexical competition.

Talker Variability. Unexpectedly, the effect of talker variability was not significant for any of the three conditions, as shown in Figure 4. A two-way ANOVA (response format x talker variability) was performed, collapsed across level of degradation and lexical competition. The main effect of response format was again significant ($F(2, 141) = 413.2, p < .001$). The main effect of talker variability, however, was only marginally significant ($F(1, 141) = 2.8, p = .09$), although the difference between single and multiple talker blocks was in the predicted direction. Planned post-hoc t-tests revealed a non-significant difference for all three conditions: $t(23) = .5, p = .62$ for the ‘before’ condition; $t(23) = 1.48, p = .15$ for the ‘after’ condition; $t(23) = 1.31, p = .20$ for the open-set condition. The response format x talker variability interaction was not significant.

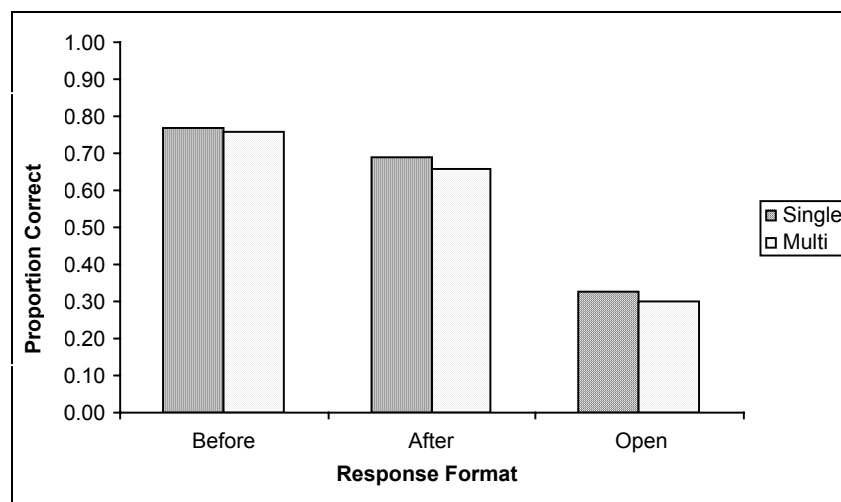


Figure 4. Word recognition performance as a function of response format and talker variability, collapsed across level of degradation and lexical competition.

Lexical Competition. The effect of lexical competition was significant only for the open-set condition, as shown in Figure 5. A two-way ANOVA (response format x lexical competition), collapsed across level of degradation and talker variability, again confirmed a significant main effect of response format ($F(2, 141) = 623.0, p < .001$). The main effect of lexical competition was not significant ($F(1, 141) = 3.4, p = .70$), but there was a significant response format x lexical competition interaction ($F(2, 1) = 4.3, p = .02$). Planned post-hoc t-tests revealed a significant effect of lexical competition only in the open-set condition: $t(23) = -.96, p = .35$ for the ‘before’ condition; $t(23) = .73, p = .48$ for the ‘after’ condition; $t(23) = 4.67, p < .001$ for the open-set condition.

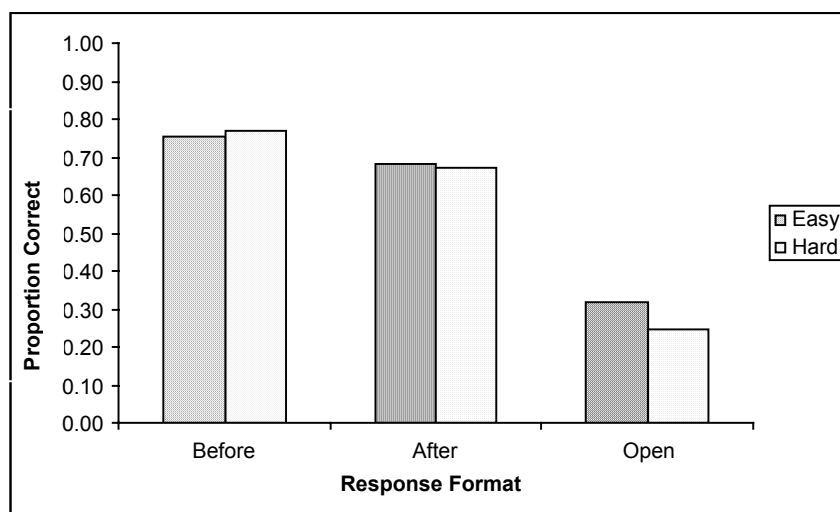


Figure 5. Word recognition performance as a function of response format and lexical competition, collapsed across level of degradation and talker variability.

Discussion

As expected, word recognition performance was affected by the signal degradation in all conditions. Unexpectedly, we did not find effects of talker variability in any of the conditions in the experiment. In previous studies, talker variability effects have been found in open-set tasks for normal-hearing listeners in noisy conditions (Sommers et al., 1997), hearing-impaired listeners in quiet conditions (Sommers et al., 1997), and normal-hearing listeners for lexically ‘easy’ and lexically ‘hard’ words (Mullennix et al., 1989). A more detailed analysis of the responses revealed that 12 of the 24 listeners in the ‘before’ condition, 14 of the 24 listeners in the ‘after’ condition, and 13 of the 24 listeners in the open-set condition performed better on the single talker blocks than the multiple talker blocks. The absence of an effect of talker variability is therefore not due to a consistent, but small, performance difference across listeners. Rather, it is due to the fact that nearly half of the participants showed an improvement in the opposite direction.

This difference between the current results and those found in the previous studies may be due to differences in experimental design. Unlike Sommers et al. (1997) who used five male and five female talkers or Mullennix et al. (1989) who used 15 different talkers, we used only four male talkers in the multi-talker blocks. It is possible that the absence of a talker effect could be due to either the small number of talkers used in the multi-talker blocks or the fact that the talkers were all male. In addition, the process of degradation using the bit-flipping algorithm may have resulted in making the talkers perceptually more similar than the degradation used in other studies, and therefore reduced the talker

variability effect. Both Sommers et al. and Mullennix et al. used white noise to reduce performance. Finally, in this experiment talker-variability was treated as a within-subjects variable. In Mullennix et al. and Sommers et al., it was treated as a between-subjects variable. Obviously, further studies investigating these response conditions are needed to confirm under which conditions talker variability effects occur and to determine whether the absence of the effect of talker variability in the present study was due to an effect of the number or gender of the talkers, the type of degradation used, or the design of the experiment itself with respect to between- and within-subject variables.

As predicted, however, the effects of lexical competition were found only in the open-set condition, confirming that even when the response alternatives are presented to listeners after a delay of one second, lexical access still occurs through a process that may be fundamentally different than the one used in open-set situations. The present set of findings confirms the need for the use of open-set tests in clinical situations and suggests that even a delayed closed-set format may not be a suitable alternative to open-set tests of speech perception and spoken word recognition. It is possible that a longer delay before the response alternatives are presented or a more complex divided attention task might reveal the same effects as an open-set task.

Conclusions

Earlier studies have reported that lexical competition effects are not observed in closed-set word recognition tasks. The present investigation replicated and confirmed those results and provided further evidence for a disparity in performance between open-set and closed-set speech intelligibility tests. In particular, even a modified closed-set test involving a delay before the presentation of the response alternatives did not reveal lexical competition effects in normal-hearing listeners under noisy conditions. Therefore, this kind of delayed closed-set format may not be any more useful in clinical situations than the traditional closed-set tasks with simultaneous presentation of the stimulus and the response alternatives. This study does, however, contribute to the growing literature on the processes involved in word recognition and suggests that higher level processes, such as lexical access that are affected by the lexical properties of the test words can be deferred in favor of other processes such as pattern-matching, even over a delay of 1000 ms.

References

- Black, J.W. (1957). Multiple-choice intelligibility tests. *Journal of Speech and Hearing Disorders*, 22, 213-235.
- Foster, J.R. & Haggard, M.P. (1987). The Four Alternative Auditory Feature test (FAAF) – Linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology*, 21, 165-174.
- Goh, W. (2001). Talker variability effects in short-term and long-term memory for spoken words. Doctoral dissertation, Indiana University.
- Horii, Y., House, A.S., & Hughes, G.W. (1970). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, 49, 1849-1856.
- House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-set response. *Journal of the Acoustical Society of America*, 37, 158-166.
- IEEE. (1969). IEEE recommended practice for speech quality measurements. (IEEE Report No. 297).
- Juszyk, P.W. & Luce, P.A. (2002). Speech perception and spoken word recognition: Past and present. *Ear & Hearing*, 23, 2-40.

- Kirk, K.I., Pisoni, D.B., & Osberger, M.J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear & Hearing, 16*, 470-481.
- Kučera, F. & Francis, W. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.
- Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear & Hearing, 19*, 1-36.
- Meyer, T.A. & Pisoni, D.B. (1999). Some computational analyses of the PBK test: Effects of frequency and lexical density on spoken word recognition. *Ear & Hearing, 20*, 363-371.
- Miller, G.A. (1946a). Articulation testing methods. In *Transmission and Reception of Sounds under Combat Conditions*. Summary Technical Report of Division 7, NDRC. Washington, D.C. pp. 69-80.
- Miller, G.A. (1946b). Intelligibility of speech: Special vocabularies. In *Transmission and Reception of Sounds under Combat Conditions*. Summary Technical Report of Division 7, NDRC. Washington, D.C. pp. 81-85.
- Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365-378.
- Nakanishi, Y. (1988). Competing noise for the evaluation of hearing-aid performance: Comparison of multi-talker babble vs. speech-shaped noise. *RIEEC Report, 37*, 9-20.
- Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report No. 10* (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Owens, E., Kessler, D.K., Telleen, C.C., & Schubert, E.D. (1981). The Minimal Auditory Capabilities (MAC) Battery. *Hearing Aid Journal, 34*, 9-34.
- Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31*, 273-279.
- Reicher, G.M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 275-280.
- Saldana, H.M., Pisoni, D.B., Fellowes, J.M., & Remez, R.E. (1996). Audio-visual speech perception without speech cues. *Proceedings of the International Conference on Speech and Language Processing, 96*, 2187-2190.
- Sommers, M.S., Kirk, K.I., & Pisoni, D.B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners I: The effects of response format. *Ear & Hearing, 18*, 89-99.
- Wheeler, D.D. (1970). Processes in word recognition. *Cognitive Psychology, 1*, 59-85.

Appendix A

Easy Words

Target	Foils				
beach	teach	reach	batch	beef	beat
book	shook	took	back	bull	bush
bud	thud	dud	bid	bum	buck
buff	puff	cuff	beef	bug	bum
cave	pave	shave	curve	cage	cane
check	deck	neck	choke	chuck	chair
coil	soil	boil	curl	coal	coin
cub	pub	tub	cob	cuff	cup
cud	mud	dud	could	cuff	cub
cuff	muff	puff	calf	cub	cut
dab	gab	jab	dub	dash	dad
death	den	deaf	debt	deck	dead
deck	neck	check	dock	debt	den
dike	bike	tyke	duke	dine	dial
dim	rim	him	dome	dish	did
dub	rub	pub	dab	dud	does
dug	thug	chug	dog	dutch	does
fang	pang	tang	gang	fad	fat
fern	churn	yearn	faun	fan	firm
fib	rib	bib	fizz	fish	fit
fig	jig	big	fog	fill	fit
gang	hang	rang	gong	gag	gas
heap	sheep	leap	hop	heal	heat
hire	tire	fire	hair	hike	hive
hive	chive	dive	heave	hike	hire
hook	cook	shook	hike	hoof	hood
hop	chop	mop	hoop	hip	hot
jam	lamb	dam	germ	jab	jack
job	rob	bob	jab	jot	jar
kid	hid	did	cod	kill	king
king	ring	thing	kick	kid	kill
lush	gush	mush	lash	luck	love
map	gap	nap	mop	mad	man
moose	deuce	noose	mace	mood	moon
mop	hop	chop	mope	mock	mob
nab	gab	jab	knob	nag	knack
neck	check	wreck	nick	knock	net
pass	gas	mass	peace	pal	pat
path	math	bath	patch	pat	pass
peach	reach	leach	perch	peal	peak
peas	seize	these	poise	peal	peat
perk	jerk	work	puck	pearl	perch
pig	rig	jig	peg	pin	pitch
pub	dub	cub	rub	pun	pup

puff	tough	cuff	pub	puck	putt
pup	cup	pip	peep	puck	pub
pus	bus	fuss	pass	pub	pun
rang	gang	sang	ring	ran	wrap
rib	fib	bib	rob	ring	rick
rise	size	guise	rose	rhyme	right
rouse	rout	raise	rose	rise	dowse
rub	tub	pub	robe	rum	rush
sag	lag	gag	sash	sack	sad
sang	hang	rang	sung	sack	sat
save	cave	gave	serve	sail	safe
shop	hop	top	ship	shock	shot
sob	cob	lob	sub	sod	psalm
soil	boil	royal	sail	sole	cell
sub	pub	rub	sob	suck	some
tab	gab	nab	tube	tang	tag
tang	pang	sang	tongue	tack	tab
team	theme	seam	term	tease	teeth
tease	cheese	peas	these	teach	team
top	hop	shop	tape	type	tar
turf	serf	tough	term	tern	Turk
wedge	hedge	ledge	wage	well	web

Hard Words

Target	Foils				
bait	date	gate	bite	beige	bathe
ban	man	fan	bean	bash	bang
bead	weed	knead	bed	beak	bean
beak	leak	reek	balk	bean	bead
bean	scene	lean	ban	beak	bead
bed	dead	red	bad	bell	bet
bill	pill	will	ball	bit	big
bought	sought	taught	bite	ball	boss
bun	pun	shun	burn	bug	bud
cake	fake	wake	coke	cane	came
den	ken	hen	dine	deaf	dare
dill	mill	hill	doll	dish	dip
dud	mud	thud	dude	dug	dub
fate	date	mate	foot	fail	fame
feel	kneel	deal	fail	full	feat
fill	dill	hill	fail	fizz	fig
fun	sun	gun	phone	fuss	fudge
gill	fill	sill	gull	ghoul	give
heal	feel	meal	hail	heed	heath
hid	lid	mid	hood	hip	hick
hip	chip	zip	heap	hill	hick
hot	lot	got	hit	hut	hop
kin	fin	shin	keen	kick	kit

kit	mitt	bit	cot	kin	kill
lace	mace	chase	lease	lame	lake
lame	maim	tame	limb	lane	lake
led	wed	red	load	let	less
mad	bad	pad	mood	mash	man
male	pail	hail	mull	main	make
mat	sat	rat	moat	math	mad
meat	seat	feat	might	meal	mean
muck	tuck	shuck	make	mull	mutt
neat	heat	peat	gnat	knead	niece
nut	shut	rut	gnat	null	none
pace	face	vase	pass	pail	pave
pale	wail	tale	pool	pain	page
pan	fan	tan	pin	patch	pat
peat	feat	seat	pout	peas	peep
pot	shot	lot	peat	palm	pop
puck	shuck	chuck	perk	putt	pub
pun	bun	done	pawn	puff	puck
race	base	face	rice	rain	rate
rake	bake	take	wreck	raid	rave
rat	gnat	hat	rut	rap	ran
rave	pave	shave	cave	rage	rate
ride	hide	tide	road	writhe	right
sack	hack	knack	soak	sad	sat
sad	mad	bad	side	sang	sap
sale	mail	vail	soil	save	sage
sane	cane	lane	sin	sake	sage
sap	yap	gap	sip	sash	sad
seed	weed	deed	side	seize	seat
seek	weak	beak	soak	seem	seat
seep	peep	reap	sit	siege	seat
sit	fit	hit	set	sill	sing
sun	done	one	sign	sub	such
tale	pail	mail	pool	tape	tame
tan	can	man	tune	tag	tang
teak	leak	beak	tyke	teach	teeth
teal	veal	real	tall	tease	teach
tier	leer	veer	tar	tease	teach
tot	jot	cot	tote	tight	tar
wed	fed	dead	wade	wedge	web
wick	hick	pick	work	witch	wig
wig	pig	dig	wag	witch	wick
wit	sit	mitt	watt	wing	with