Word Recognition in Multitalker Babble Measured with Two Psychophysical Methods

Richard H. Wilson*†
Christopher A. Burks*
Deborah G. Weakley*

Abstract

The purpose of this experiment was to determine the relationship between psychometric functions for words presented in multitalker babble using a descending presentation level protocol and a random presentation level protocol. Forty veterans (mean = 63.5 years) with mild-to-moderate sensorineural hearing losses were enrolled. Seventy of the Northwestern University Auditory Test No. 6 words spoken by the VA female speaker were presented at seven signal-to-babble ratios from 24 to 0 dB (10 words/step). Although the random procedure required 69 sec longer to administer than the descending protocol, there was no significant difference between the results obtained with the two psychophysical methods. There was almost no relation between the perceived ability of the listeners to understand speech in background noise and their measured ability to understand speech in multitalker babble. Likewise, there was a tenuous relation between pure-tone thresholds and performance on the words in babble and between recognition performance in quiet and performance on the words in babble.

Key Words: Hearing loss, multitalker babble, signal-to-babble ratio, speech perception, word-recognition performance

Abbreviations: S/B = signal-to-babble
Different psychophysical procedures can produce different results. An example in the audiology literature is the Békésy Ascending Descending Gap Evaluation (BADGE) described by Hood et al (1964). In that study a threshold several decibels lower was observed when the search for threshold started at a level above the estimated threshold (descending) as opposed to when the search for threshold started at a level below the estimated threshold (ascending). With the descending method, the listener is aware of the signal and its characteristics at levels substantially above the level of minimum audibility. In contrast to the descending procedure, with the ascending method the listener is not aware of the signal and its characteristics (i.e., the listener is uncertain about the characteristics of the target signal) until the signal is above the level of minimum audibility.

With speech stimuli, three psychophysical methods have been used to study speech-recognition thresholds. First, the commonly used bracketing or “up-down” method (Newby, 1958; Jerger et al, 1959) closely follows the Hughson-Westlake protocol for pure-tone thresholds (Carhart and Jerger, 1959). Second, the descending method used with PAL Auditory Test No. 9 (Hudgins et al, 1947) and CID W-2 (Hirsh et al, 1952) involves a select number of words presented at each of a sequence of presentation level decrements. Third, Chaiklin et al (1967) advocated an ascending approach for use with patients having suspected pseudohypacusis. Jahner et al (1994) compared thresholds for spondaic words using ascending and descending protocols. As with the pure-tone protocols described earlier, the thresholds for the descending protocol (25.7 dB HL) were lower than thresholds for the ascending protocol (29.0 dB HL). To our knowledge, no comparisons of speech-recognition performances have been made for ascending or descending protocols versus a randomized presentation protocol.

Probably the most common complaint that older listeners have about their hearing is their inability to understand speech in a background noise. Clinically, this complaint is seldom addressed in the course of an audiological evaluation. In an attempt to devise a protocol that audiologists would use to assess the ability of the patient to understand speech in a background noise, a word-recognition paradigm in multitalker babble was developed, the results from which make a clear distinction in terms of signal-to-noise (babble) ratio hearing loss between the performances by listeners with normal hearing and listeners with hearing loss (Wilson, 2003). The primary purpose of this current experiment was to determine the relationship between psychometric functions for monosyllabic words presented in multitalker babble using a descending presentation level method and a random presentation level method. With the descending presentation level paradigm, all words were presented at the most favorable signal-to-babble ratio (S/B) before the words were presented sequentially at the lower signal-to-babble ratios. With the random presentation level paradigm, all of the words were presented at random signal-to-babble ratios. With the descending protocol, when all words at a particular signal-to-babble ratio were incorrect, the stopping rule terminated the procedure. Typically, then, with the descending paradigm, only a portion of the test materials was required because the stopping rule precluded presentation of the test materials at the lowest signal-to-babble ratios. In contrast to the descending protocol, the random protocol required that all of the test materials be presented before the test was terminated.

The question concerning the possible difference between the data from the two procedures was precipitated for clinical reasons, that is to say, the time required to administer the procedure that is critical in a clinic setting. It was anticipated that the psychometric function for the random presentation level procedure would be displaced slightly to the higher signal-to-babble ratios with a more gradual slope than
the function for the descending presentation level procedure. This relation was expected because of the uncertainty associated with the randomized procedure; that is, in the random procedure from one word presentation to the next word presentation; the listener was uncertain of the signal-to-babble ratio of the presentation. The more uncertainty there is associated with a listening condition, the more difficult the listening/response task is to the listener and the more gradual the slope of the psychometric function.

Study of the effects of uncertainty by the listener about characteristics of the signal has a history in the signal detection literature. Green and Swets (1966) report a simple example in which a 95% level of performance was achieved when the frequencies (500 and 1000 Hz) were presented separately, that is, the stimulus frequency from trial to trial was the same. When the listener did not know from trial to trial which of the two signals would be presented, which increased listener uncertainty, the level of performance dropped to about 73% on both signals. Recently, in a study of informational masking, Lutfi et al (2003) compared the psychometric functions for pure-tone signals masked by broadband noise (low uncertainty) and by multitone maskers whose frequencies and amplitudes varied from presentation to presentation (high uncertainty). Lutfi et al observed the more uncertain the masker, the more gradual the slope of the psychometric function.

A secondary purpose of this study was to examine the relation between the perceived ability of the listener to understand speech in background noise and the objective measure of their ability to understand speech in background noise. In a pilot study, the 16 questions from the background noise subscale of the Profile of Hearing Aid Benefit (PHAB; Cox and Gilmore, 1990) were used to establish the subjective measure of the ability of the listener to understand speech in noise. Because of the confusion created by the wording of some of the PHAB subscale questions and the time required to administer the PHAB, the subscale of the PHAB was abandoned in favor of two simple questions (see Appendix A). The questions involved asking the listeners to rate on a scale of 1 to 10 (1 being no difficulty and 10 being extreme difficulty) their ability to understand words in quiet and in noise. No other information was volunteered about the circumstances of either of the two listening environments. Finally, an examination was made of the relations among the pure-tone thresholds and the word-recognition performance in multitalker babble.

**METHODS**

**Materials**

The development of the words in multitalker babble materials is described in Wilson (2003) and Wilson et al (2003). The materials consist of 70 monosyllabic words from Northwestern University Auditory Test No. 6 (NU-6; Tillman and Carhart, 1966) spoken by a female speaker (Department of Veterans Affairs, 1998). Based on recognition performance data in which each word was presented time locked in a unique segment of babble, ten words were selected for presentation at each of seven signal-to-babble ratios from 0 to 24 dB in 4 dB steps (i.e., the ten words presented at 24 dB S/B were always presented at that level, with ten different words always presented at 20 dB S/B, etc.). For the current study, two randomizations of the words in the descending presentation level paradigm and two randomizations of the words in the random presentation level paradigm were prepared by concatenating the word/babble segments as appropriate using in-house, batch-file routines. Because the segments were concatenated at the negative going zero crossings, the boundaries between word/babble segments were acoustically and perceptually transparent to the listener. The materials with the speech and babble mixed were recorded on a compact disc (Hewlett-Packard, Model DVD200i). The carrier phrases and words each were about 1.2 sec, the interstimulus intervals were about 2.7 sec, and the total track times including miscellaneous intervals were 4 min. and 42 sec (4:42).

**Subjects**

Forty older, male listeners (mean age = 63.5 years) with sensorineural hearing loss participated in the study. The listeners with hearing loss met the following inclusion criteria: (1) threshold at 500 Hz \( \leq 30 \) dB HL (ANSI, 1996), and (2) threshold at 1000 \( \leq 40 \)
Word Recognition in Multitalker Babble/Wilson et al

For the descending presentation level procedure, the ten words at 24 dB S/B were randomized, followed by the ten words at 20 dB S/B, etc. For the random presentation level procedure, all of the 70 words at their respective signal-to-babble ratios were randomized. The words in babble were presented monaurally with the level of the babble fixed at 80 dB SPL, and the level of the words varied from 80 to 104 dB SPL. The materials were reproduced on a CD player (Sony, Model CDP-497), fed through an audiometer (Grason-Stadler, Model 61), and delivered to a TDH-50P earphone encased in a Telephonics P/N 510C017-1 cushion. All testing was conducted in a sound booth, and the verbal responses of the listeners were recorded into a spreadsheet. Data collection took about 15 minutes.

RESULTS

Two metrics were compiled on the recognition performances of each listener under the descending and random presentation paradigms. First, the Spearman-Kärber equation (Finney, 1952; Wilson et al, 1973) was used to calculate the 50% point on each of the 80 psychometric functions (40

Table 1. The Mean Percent Correct for the Descending and Random Presentation Orders at Each of the Seven Presentation Levels (dB S/B)

<table>
<thead>
<tr>
<th>dB S/B</th>
<th>DESCENDING PRESENTATION</th>
<th>RANDOM PRESENTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD (WORDS)</td>
</tr>
<tr>
<td>24</td>
<td>90.3</td>
<td>9.3</td>
</tr>
<tr>
<td>20</td>
<td>85.8</td>
<td>5.1</td>
</tr>
<tr>
<td>16</td>
<td>69.3</td>
<td>8.7</td>
</tr>
<tr>
<td>12</td>
<td>54.0</td>
<td>12.3</td>
</tr>
<tr>
<td>8</td>
<td>19.8</td>
<td>8.9</td>
</tr>
<tr>
<td>4</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Note: The corresponding standard deviations for the words at each level and for the subjects at each also are listed.
listeners by two conditions). Second, the percent correct at each presentation level for each listener in the two presentation orders was compiled. A nested analysis-of-variance (ANOVA) then was used to examine the data at the 50% point obtained with the Spearman-Kärber equation. There were no significant differences between the data from the order of presentation, this is, which procedure was presented first or second (F[1,32] = 0.309, p = 0.582). Additionally, no significant difference was found between the two randomizations of each presentation level paradigm (the nested variables) (descending, F[2,32] = 0.291, p = 0.749; random, F[2,32] = 0.204, p = 0.817). Because there were no significant differences among these variables, the results that follow are for the combined data for each of the two presentation level paradigms.

The percent correct recognition at the seven signal-to-babble ratios for the descending (triangles) and random (circles) presentation level paradigms are presented in Figure 2 and are listed in Table 1. Table 1 also includes the standard deviations for the words and subjects at each level. The lines through the datum points in Figure 2 are the best-fit, third-degree polynomials used to describe the data. The 50% correct points and the slopes of the functions at the 50% correct points in Figure 2 are listed in Table 2. As seen from the figure and tables, the functions for the descending and random conditions are separated by 0.3 dB with identical slopes (6.5%/dB). The 50% points calculated with the Spearman-Kärber equations are in good agreement (within 0.6 to 0.7 dB) with the 50% points calculated

**Table 2. The Mean 50% Correct Points (dB S/B) and Standard Deviations (dB) Established with the Spearman-Kärber Equation and the 50% Correct Points (dB S/B) and Slopes at the 50% Points (%/dB) Calculated with the Polynomial Equations Used to Fit the Mean Data in Figure 2**

<table>
<thead>
<tr>
<th></th>
<th>Descending Presentation</th>
<th>Random Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman-Kärber</td>
<td>Random</td>
</tr>
<tr>
<td>Mean</td>
<td>SD</td>
<td>50%</td>
</tr>
<tr>
<td>13.1</td>
<td>4.0</td>
<td>12.4</td>
</tr>
</tbody>
</table>

**Figure 2.** The mean percent correct recognition at the respective signal-to-babble ratios obtained from the 40 listeners under the random (circles) and descending (triangles) presentation level paradigms in multitalker babble. The lines through the functions are the best-fit, third-degree polynomials (random, \( y = 0.6524 - 1.2356x + 0.6799x^2 - 0.0198x^3 \); descending, \( y = -1.3548 - 1.1925x + 0.6670x^2 - 0.0192x^3 \)). The thick line is for the random condition, and the thin line is for the descending condition.
from the polynomial equations. The standard deviations in Table 1 indicate substantially more variability for the subjects than for the words with the standard deviations for both presentation paradigms about the same. As anticipated, larger standard deviations were associated with the dynamic portions of the psychometric function. Because of their lack of homogeneity, the subjects were expected to have more variability than the words, which were relatively homogeneous because they were selected based on their homogeneity at given signal-to-babble ratios (Wilson, 2003).

Figure 3 is a bivariate plot of the percent correct recognition for each of the 70 words presented in the descending method (abscissa) and in the random method (ordinate). The filled symbol illustrates the mean data. The solid diagonal line represents equal performance on the two methods, whereas the dashed line is a linear regression fit to the data ($r^2 = 0.97$). Because of superimposed datum points in the figure, the numbers in parentheses are included to represent the number of datum points above, on, and below the diagonal line. Although performance on 32 of the words was better with the random method than with the descending method, it is impressive that the slope of the regression is 0.97%/%, which indicates little difference between the performances on the materials presented with the two protocols. The nested ANOVA indicated no significant difference between the 50% points on the functions for the descending and randomized psychophysical methods ($F(1,32) = 3.259, p = 0.080$). Thus, both the graphic analysis and inferential statistics indicate that the descending and random presentation level protocols produce the same results.

Finally, each listener was asked to rate on a scale of 1 to 10 (10 being the most difficult) how much difficulty he had understanding speech in quiet and in background noise. All listeners expressed that they had more difficulty understanding speech in background noise than in a quiet background. The mean responses were 3.6 in quiet and 7.6 in noise with 2.0 and 1.8 standard deviations, respectively. Assuming a linear psychological scale, the patients indicated that listening in noise was about twice as difficult as listening in quiet. The relationship between the subjective rating of each listener to understand speech in noise and the objective value obtained on their listening in background noise was
examined and is illustrated in Figure 4 with the response to the noise question on the abscissa and the decibel signal-to-babble ratio at which the 50% point was measured for both the descending (triangles) and random (circles) conditions on the ordinate. The two lines in the graph are linear regressions fit to the data for the respective conditions. For both presentation methods, the slopes of the lines approached zero, indicating a minimum relation between the objective and subjective variables. Although the listeners perceived that they had more trouble understanding speech in background noise than in quiet, the relation between their perceived difficulty to understand speech in background noise and the measured difficulty was lacking for the most part. This finding is consistent with other reports that minimize the relation between objective and subjective measures of the difficulty listeners with hearing loss have (and report that they have) understanding speech in background noise (Rowland et al, 1985).

Finally, a main issue that prompted this study was the time required to administer the test sequence. The complete descending and random protocols each required 4:42 (282 sec) for 70 words. The stopping rule with the descending method terminated any presentations following a level at which no correct responses were obtained. Thus, if a listener incorrectly responded to all stimuli at 4 dB S/B, then the stimuli at 0 dB S/B were not given, and the total test time was 240 sec (4:00). With the descending method, 2, 17, 14, 6, and 1 listener(s) required 70, 60, 50, 40, and 30 words, respectively, to complete the protocol. The mean test time for the descending protocol was 3:33 (53.3 words) with a standard deviation of 0:37 (8.9 words). Thus, the average test time for the descending protocol was 1:09 shorter than the 4:42 test time required for the random protocol.

**DISCUSSION**

The 50% correct points for the individual functions ranged from 5 to 16 dB S/B with means of 12 to 13 dB S/B (Table 2) and are in excellent agreement with previous data from our lab on the words in multitalker babble task with listeners having similar hearing losses (Wilson 2003; Wilson et al, 2003). For young adults with normal hearing, the 90th percentile for the 50% correct points for the words in multitalker babble is 6 dB S/B (Wilson 2003; Wilson et al, 2003). In the current study, only one listener with hearing loss was within this normal range. The 62-year-old listener had only a mild high-frequency hearing loss with a three-frequency, pure-tone average of 13.3 dB HL. Although there is only a 5–6 dB mean difference between listeners with normal hearing and listeners with hearing loss, we must remember that this difference is in terms of signal-to-noise ratio. As indicated by many investigators (e.g., Carhart and Tillman, 1970; Plomp, 1978; Duquesnoy and Plomp, 1983; Dubno et al, 1984; Wilson et al, 1996; Killion, 2002), a hearing loss in terms of signal-to-noise ratio is devastating and is the difficult component of hearing loss to rehabilitate.

The relationships between recognition performance on the words in babble and other auditory tasks like pure-tone sensitivity are tenuous at best. Figure 5 is a series of
bivariate plots in which the 50% correct points for the words in babble in the descending presentation paradigm (ordinate) are plotted as a function of the pure-tone threshold (in dB HL) (abscissa). In each panel, the line represents the linear regression fit to the data, and the filled symbol represents the mean datum points. For the individual frequencies (1000–8000 Hz), the slopes of the lines were 0.05 to 0.06 dB/dB. Thus, for every decibel change in the pure-tone threshold there was a corresponding 0.05–0.06 dB change in the level at which the 50% correct point was for the words in babble task. The data in Figure 6 depict a similar relation between word-recognition performance in quiet and in multitalker babble. The slope of the linear regression in Figure 6 is -0.06 dB/%, which means that as word recognition in quiet increased by 1%, there was only a corresponding 0.06 dB decrease in the level of the 50% correct point in babble. These relations indicate that word-recognition performance in multitalker babble cannot be predicted (within general limits) from performances on other auditory tasks. The relation that we can predict, however, is that poor word-recognition abilities in quiet certainly predict poor word-recognition abilities in multitalker babble. As succinctly stated by Killion (2002), if you want to know the ability of a listener to recognize speech in background noise, then you have to measure that ability.

Initially we anticipated that performance on the random presentation level paradigm would be slightly poorer than performance on the descending method. This relation was expected because of the uncertainty thought to be associated with the subject not knowing what levels to listen “for” in the random paradigm, that is, where on the signal-to-babble ratio continuum to listen. No difference in performance between the two methods, however, was observed. Introspective reports from the listeners indicated that they had to “try harder” on the random method, which the subjects perceived as a more difficult listening condition than the descending method. Perhaps these reports from the subjects indicate that their approach to listening in the random paradigm was not one of uncertainty regarding the levels of the words, but rather, their approach was along the lines of “worse case scenario,” that is, not knowing the signal-to-babble ratio of the next word presentation, the subjects always were prepared to listen for words that would be presented at the poorer signal-to-babble ratios.

In summary, although the descending protocol required less time (69 sec) to administer than the random protocol, there was no significant difference between the word-recognition abilities in multitalker babble obtained with the two paradigms. As has been demonstrated previously, older listeners with hearing loss for pure tones typically have a 5+ dB hearing loss in terms of signal-to-noise ratio, which is an auditory deficit that cannot be predicted from other measures of auditory behavior. This signal-to-noise ratio hearing loss is the component of generalized hearing loss that is most often mentioned as a patient complaint and is the most difficult to rehabilitate.

REFERENCES


APPENDIX A

On a scale of 1 to 10, please answer the following two questions.

1. When listening to a conversation in quiet without your hearing aids, how difficult is it for you to understand what the speaker is saying?

On the scale of 1 to 10, 1 means no difficulty understanding and 10 means extreme difficulty.

2. When listening to a conversation in a noisy background without your hearing aids, how difficult is it for you to understand what the speaker is saying?

On the scale of 1 to 10, 1 means no difficulty understanding and 10 means extreme difficulty.

Note: In both questions, “without your hearing aids” is eliminated if the listener does not have hearing aids.