Published in final edited form as:
Ear Hear. 2012 ; 33(1): 112–117. doi:10.1097/AUD.0b013e31822c2549.

Development and validation of the AzBio sentence lists

Anthony J. Spahr, Ph.D.,
Arizona State University

Michael F. Dorman, Ph.D.,
Arizona State University

Leonid M. Litvak, Ph.D.,
Advanced Bionics Corporation

Susan Van Wie, M.S.,
Arizona State University

Rene H. Gifford, Ph.D.,
Vanderbilt University

Philipos C. Loizou, Ph.D.,
University of Texas at Dallas

Louise M. Loiselle, M.S.,
Arizona State University

Tyler Oakes, Au.D., and
Arizona State University

Sarah Cook, B.S.
Arizona State University

Abstract

Objectives—This goal of this study was to create and validate a new set of sentence lists that could be used to evaluate the speech perception abilities of hearing impaired listeners and cochlear implant users. Our intention was to generate a large number of sentence lists with an equivalent level of difficulty for the evaluation of performance over time and across conditions.

Design—The AzBio sentence corpus includes 1000 sentences recorded from 2 female and 2 male talkers. The mean intelligibility of each sentence was estimated by processing each sentence through a 5-channel cochlear implant simulation and calculating the mean percent correct score achieved by 15 normal-hearing listeners. Sentences from each talker were sorted by percent correct score and 165 sentences were selected from each talker and were then sequentially assigned to 33 lists, each containing 20 sentences (5 sentences from each talker). List equivalency was validated by presenting all lists, in random order, to 15 cochlear implant users.

Correspondence. Anthony J. Spahr, Department of Speech and Hearing Science, Arizona State University, Lattie F. Coor Hall, Room 3462, Tempe, Arizona 85287-0102, Phone: (480)965-8167, Fax: (480)965-8516, tspahr@asu.edu.

Results—Using sentence scores from the cochlear implant simulation study produced 33 lists of sentences with a mean score of 85% correct. The results of the validation study with cochlear implant users revealed no significant differences in percent correct scores for 29 of the 33 sentence lists. However, individual listeners demonstrated considerable variability in performance on the 29 lists. The binomial distribution model was used to account for the inherent variability observed in the lists. This model was also used to generate 95% confidence intervals for one and two list comparisons. A retrospective analysis of 172 instances where research subjects had been tested on two lists within a single condition revealed that 94% of results were accurately contained within these confidence intervals.

Conclusions—The use of a 5-channel cochlear implant simulation to estimate the intelligibility of individual sentences allowed for the creation of a large number of sentence lists with an equivalent level of difficulty. The results of the validation procedure with cochlear implant users found that 29 of 33 lists allowed scores that were not statistically different. However, individual listeners demonstrated considerable variability in performance across lists. This variability was accurately described by the binomial distribution model and was used to estimate the magnitude of change required to achieve statistical significance when comparing scores from one and two lists per condition. Fifteen sentence lists have been included in the AzBio Sentence Test, for use in the clinical evaluation of hearing impaired listeners and cochlear implant users. An additional 8 sentence lists have been included in the Minimum Speech Test Battery to be distributed by the cochlear implant manufacturers for the evaluation of cochlear implant candidates.

Introduction

Gifford, Shallop, and Peterson (2008) evaluated the performance of hearing aid users and cochlear implant users on new and traditional tests of speech recognition. They reported that a new set of materials, the AzBio sentences, produced results that were highly correlated with monosyllabic word scores and did not suffer the same ceiling effects in quiet as other sentence materials. For these reasons, Gifford et al. (2008) suggested that the AzBio sentences could be of value in the clinical evaluation of adult hearing-impaired listeners and cochlear implant users. This suggestion was later echoed by a committee of audiologist clinician/scientists who recommended the use of the AzBio sentence metric for both pre- and post-implant assessment of sentence recognition performance (Fabry et al., 2009). Since that time, the cochlear implant manufacturers in the United States have moved to include AzBio sentence lists in a new battery of tests that will serve as the standard for evaluation of pre- and post-implant assessments of speech recognition. The potential for widespread use of these materials has prompted this report describing the development and validation of the current AzBio sentence lists.

The AzBio sentences, first described in Spahr and Dorman (2004), were developed in the Department of Speech and Hearing Science at Arizona State University. The sentences were created specifically for an experiment (Spahr and Dorman, 2005; Spahr, Dorman and Loiselle, 2007) comparing the speech understanding abilities of high-performing patients implanted with different cochlear implant systems (Advanced Bionics Corporation; Cochlear Corporation; Med El Corporation). The goals for these materials were to (i) provide an unbiased evaluation of individuals with extensive exposure to traditional sentence materials, (ii) allow for evaluation of performance in a large number of test conditions, (iii) create lists of sentences with similar levels of difficulty for within-subject comparisons, and (iv) provide an estimate of performance that was consistent with the patient's perception of their performance in everyday listening environments. The development of these sentence materials was enabled by a grant from the Arizona Biomedical Institute at Arizona State University (currently known as the Biodesign Institute). In appreciation, the resulting speech materials were dubbed the AzBio sentences.

Methods

Sentence Construction

In total, 1500 sentences were created for the AzBio corpus. Sentence length was limited to between 3 and 12 words (mean = 7.0, s.d. = 1.4) and proper nouns were generally avoided. No other restrictions were placed on complexity, vocabulary, or phonemic content. The sentences include up-to-date, adult topics and current social ideas.

Sentence Recordings

Of the original 1500 sentences, only the final 1000 submissions were recorded for possible inclusion in the AzBio corpus. Four talkers, two male (ages 32 and 56) and two female (ages 28 and 30) were selected to each record 250 sentences.

During recording, talkers were seated in a sound-treated booth. The 250 sentences were recorded in blocks of 50 using an AKG C2000B condenser microphone connected to an M-Audio Audiophile USB soundcard connected to a Sony laptop computer running Cool Edit 2000 software. All recordings were made with a sample frequency of 22050 Hz and 16-bit resolution.

The microphone was placed in a boom and positioned approximately 6 – 12 inches from the talker. Each talker was instructed to speak at a normal conversational pace and volume and to avoid using overly enunciated speech. Sentence production was monitored by an examiner. In the event of mispronunciations, misread words, or any other unintended disruptions, the talker was prompted to repeat the sentence. The final production of each sentence was isolated from the recorded block and saved as a unique sound file. A global adjustment was made to the 250 recorded sentences of each talker (e.g. all recordings from a single talker were attenuated by 2 dB) to control for slight differences in recording levels across talkers. Across talkers, the average speaking rate ranged from 4.4 to 5.1 syllables per second, consistent with normal speaking rates (Goldman, 1968) and the RMS level of individual sentences had a standard deviation of 1.5 dB and a range of 9.6 dB.

Sentence Intelligibility Estimation

All 1000 sentence files were processed through a five-channel cochlear implant simulation (Dorman et al., 1998) and presented to 15 normal-hearing listeners. Listeners were seated comfortably in a sound-treated booth, instructed to repeat each sentence, and to guess when unsure about any word. Sentences were presented at a comfortable level using Sennheiser HD 20 Linear II headphones. Each listener completed a practice session with 50 TIMIT sentences (Seneff and Zue, 1988) processed through the same simulation prior to hearing the 1000 test sentences presented in random order. Each sentence was scored as the number of words correctly repeated by each listener. The mean percent correct score for each sentence (total words repeated correctly / total words presented) was used as the estimate of intelligibility. Sentence scores from each talker ranged from <20% to 100% correct.

Sentence Selection and List Formation

A pilot study evaluating our method for generating equivalent lists revealed that a minimum of 20 sentences was necessary to significantly reduce list variability. Thus, it was decided that each list would consist of 5 sentences from each of the 4 talkers and that the average level of intelligibility for each talker would be held constant across lists. For each talker, the 250 sentences were rank ordered by mean percent correct scores and a block of 165 consecutively ordered sentences was selected to create 33 lists. Still rank ordered by mean percent correct scores, the sentences from each talker were then sequentially assigned to lists, with the first 33 sentences assigned, in order, to lists 1–33 and the next 33 sentences assigned, in order, to lists 33-1 (e.g. 1, 2, 3, …3, 2, 1). This sentence-to-list assignment produced 33 lists of 20 sentences with a mean score of 85 percent correct (s.d. = 0.5). Individual sentence scores and mean list scores are shown in Figure 1. Average intelligibility of individual talkers across lists was 90.4% (s.d. = 0.5) and 86.0% (s.d. = 0.5) for the two female talkers and 87.0% (s.d. = 0.4) and 77.2% (s.d. = 0.7) for the two male talkers. Lists had an average of 142 words (s.d. = 6.4, range = 133 to 159).

List Equivalency Validation

Validation of the equivalency and inherent variability in the newly formed sentence lists was accomplished by testing 15 cochlear implant users on all 33 sentence lists. Participants had monosyllabic word scores of 36 to 88 percent correct (avg = 61%, s.d. = 16). To avoid ceiling effects, sentence lists were presented in +5 dB SNR (multi-talker noise) for subjects with word scores of 85% or greater, +10 dB SNR for subjects with word scores between 65% and 84%, and in quiet for subjects with word scores below 65%. Sentence list order was randomized for each subject and lists were tested in 5 blocks, each containing 7 lists. For each subject, the final list of block 1 was repeated as the final list of blocks 3 and 5, resulting in a total of 35 test lists. Only the score from the first presentation of each list was considered in the validation analysis.

During testing, subjects were seated in a sound-treated booth. Sentences were presented at 60 dB SPL in the sound field from a single loudspeaker at 0 degrees azimuth on the horizontal axis. Subjects were instructed to repeat back each sentence and to guess when unsure of any word. Prior to testing, subjects completed a practice list of 50 sentences that were not included in the 33 lists. Following completion of each block of sentence lists, subjects were asked to exit the sound booth and relax for a minimum of 15 minutes. Each sentence was scored as the number of words repeated correctly and a percent correct score was calculated for each list.

Results

Validation Study

The mean level of performance achieved by individual CI listeners ranged from 46 to 86 percent correct (mean = 69%, s.d. = 13). The distribution of list scores for all 15 CI listeners is shown in Figure 2. Averaged scores for the 33 sentence lists ranged from 62 to 79 percent correct (mean = 69%, s.d. = 3.8). Averaged scores for lists tested in blocks 1–5 were 68, 69, 69, 70, and 71 percent correct, respectively. Thus, there was no significant effect of practice.

The individual results of these 15 CI listeners were used to identify lists that were not of equal difficulty. Because of the range of performance levels across listeners, the list scores observed for each listener were normalized for comparison. For each CI listener, each of the 33 list scores was subtracted from that listener's mean score. This transform retains the distribution characteristics of the original list scores, but normalizes the mean score for each listener to zero. The distribution of normalized scores for all 33 lists is shown in Figure 3.

A repeated-measures ANOVA revealed a significant (alpha = 0.05) main effect of list number. A post-hoc Tukey test revealed that 4 of the 33 lists were significantly different from at least one other list. Based on this statistical analysis, lists identified as significantly easier (15, 23, and 33) or more difficult (28) than the other lists were removed from the set. The remaining 29 lists had a mean score of 68 percent correct (s.d. = 2.6) and a range of 62 – 72 percent correct, with no statistically significant differences.

Variability of Materials

Though no statistical differences were found among 29 of the 33 lists on average, all subjects demonstrated some degree of variability across lists. As with other speech materials, this variability is expected and can be modeled. Thorton and Raffin (1978) used a binomial distribution model to predict variability in monosyllabic word tests with different numbers of items. The binomial model holds that the variability of an individual's performance is a function of both the starting level of performance and the number of independent items scored in the task. Variability is highest for mid-range performance and lowest near the upper and lower ends of the range. Variability is expected to decrease as the number of independent items is increased. Based on this model, a relatively high level of variability could be expected in this study, as scores were intentionally kept off of the ceiling by adding background noise for some listeners. Given that each list contained 20 unique sentences and lists had an average of 142 words, it was expected that the number of independent items would fall somewhere between 20 and 142. To determine the number of items that would best model the observed variability, a mean and standard deviation was calculated from the percent correct scores measured on the 29 equivalent lists, for each listener. The results were then plotted against the binomial confidence intervals predicted by different numbers of list items. Visual analysis revealed that the results of these 15 cochlear implant listeners were best fit by a 40-item model, shown in Figure 4. In that figure, the solid line indicates the average expected variance as a function of mean performance level assuming a binomial 40-item model, while the dashed lines indicate the 95% confidence intervals. Both the mean and the confidence intervals were computed by applying the bootstrap estimate to the binomial distribution. This outcome suggests that variability on this set of materials should be just slightly higher than that observed on a 50-item monosyllabic word test.

List Variability

The same binomial distribution model described above was used to predict the variability of the AzBio sentence lists when only one or two lists were tested in each condition. Reducing the number of 40-item lists included in each condition increases the expected variability and, therefore, the change in performance required to achieve statistical significance. Table 1 displays the upper and lower 95% confidence intervals as a function of starting level of performance (percent correct) when comparing scores from one or two lists per condition. Caution should be used when the reference score falls below 15 percent correct or above 85 percent correct, as the function is compressed due to floor and ceiling effects, respectively. The table reveals that for a starting score of 50 percent correct, a significant change in performance would require a change of more than 15 percentage points for a single list and 11 percentage points for two sentence lists, with a single listener.

The accuracy of this model was tested with a retrospective analysis of experimental data collected at Arizona State University, Mayo Clinic, Rochester, and Advanced Bionics. A review of recent studies identified 172 instances where subjects had been tested on 2 of the 29 equivalent AzBio lists within the same condition. These data were pulled from 66 cochlear implant listeners evaluated in several different test conditions. Figure 5 displays the percent correct score for the first (A) and second (B) list tested in each condition. Because both scores were obtained from the same listener in the same condition, differences in list scores are expected to fall within the confidence intervals of the 40-item binomial model for single list comparisons. Approximately 94% of the 172 scores fall within the 95% confidence intervals for single list comparisons. Thus, the model accurately describes the variability of the test material observed when comparing single list scores within the same test condition.

Commercial Materials

AzBio Sentence Test—Based on feedback and requests from several clinical and research test sites, it was decided that a subset of the AzBio sentence lists would be released in CD format for evaluation of hearing impaired listeners and cochlear implant users. It was determined that this subset would include 15 lists that produced the most similar average level of performance, with the least variability based on the scores of the 15 cochlear implant listeners. The selected lists (2, 3, 4, 5, 10, 11, 12, 14, 16, 17, 18, 21, 22, 24, and 26) had a mean score of 68 percent correct, with individual list scores ranging from 66 to 70 percent correct. For each subject, a normalized list score was calculated by subtracting each list score from the individual’s mean score. Normalized list scores for the 15 lists are shown in Figure 6. The distribution of normalized list scores varies slightly from that shown in Figure 1, as only the relevant lists are considered in the calculation of the mean and difference scores.

Minimum Speech Test Battery—At the request of Cochlear Americas, Advanced Bionics, and Med-El Corporation, an additional subset of lists was selected for inclusion in the Minimum Speech Test Battery. The test battery includes 8 AzBio sentence lists, 12 list-pairs from the Bamford-Kowal-Bench Sentence in Noise (BKB-SIN, Etymotic Research, 2005) test (Killion et al, 2001), and 10 CNC word lists (Peterson and Lehiste, 1962) and will be distributed to cochlear implant research centers in North America for the evaluation of cochlear implant candidates and users. It was decided that this subset would not include any of the 15 lists from the AzBio Sentence Test. Of the remaining 14 lists, the 8 lists with the most similar mean scores and the least variability, based on the scores of the 15 cochlear implant listeners, were selected for this test battery. The selected lists (1, 7, 9, 19, 20, 27, 30, and 31) had a mean score of 68 percent correct (s.d. = 2.8), with individual list scores ranging from 65 to 72 percent correct. For each subject, a normalized list score was calculated by subtracting the individual listener's mean list score from the individual sentence score. Normalized list scores for the 8 lists are shown in Figure 7. The distribution of normalized list scores varies slightly from that shown in Figure 1, as only the relevant lists are considered in the calculation of the mean and difference scores.

Conclusion

As stated in the introduction, the goals for these materials were to (i) provide an unbiased evaluation of individuals with extensive exposure to traditional sentence materials, (ii) allow for evaluation of performance in a large number of conditions, (iii) create lists of sentences with similar levels of difficulty for within-subject comparisons, and (iv) provide an estimate of performance that was consistent with the patient's perception of their performance in everyday listening environments. Of the 1500 sentences written and recorded for inclusion in the AzBio sentence corpus, 1000 were evaluated using a cochlear implant simulation, 660 were used to form 33 lists of 20 sentences, and 29 of the 33 lists were found to be of equivalent intelligibility based on the scores obtained from 15 cochlear implant listeners. Because these materials are more difficult than the HINT sentences (Nilsson, Soli, and Sullivan, 1994; Gifford et al, 2008) and likely more difficult than the CUNY sentences (Boothroyd, Hnath, Hanin, and Rabin, 1988), fewer subjects should reach the ceiling in quiet or in moderate levels of noise. With such a large set of lists, researchers can test a large number of experimental conditions. These lists should also allow clinicians to track changes in performance of individual listeners over time or across conditions with greater confidence that large differences in performance are not simply due to differences in list intelligibility. Finally, patients have frequently reported that their scores on the relatively difficult AzBio sentences are consistent with their own estimation of performance in real-world environments. For these reasons, it was determined that these materials could be used successfully to evaluate speech understanding of adult patients in the clinic and the laboratory. Thus, 15 of the 29 lists have been included in the AzBio Sentence Test and another 8 have been included in the Minimum Speech Test Battery.

Presumably, clinical acceptance of these materials with adult patients will lead to use with other populations. The materials could potentially be used to evaluate speech understanding of younger listeners, hearing impaired listeners, hearing aid users, and even normal-hearing listeners under adverse listening conditions. Thus, it should be noted that further research will be necessary to assess the reliability of these materials for these specific applications and populations.

Acknowledgments

The development of these sentence materials was enabled by a grant from the Biodesign Institute at Arizona State University. The copyright for the sentence materials is held by the Arizona Board of Regents. The licensing rights to the materials are held by Auditory Potential, LLC, which is owned by co-authors Anthony J. Spahr and Michael F. Dorman.

Funding

Initial funding for this project was provided by a grant from the AzBio Design Institute at Arizona State University.

References

Boothroyd A, Hnath-Chisolm T, Hanin L, Kishon-Rabin L. Voice fundamental frequency as an auditory supplement to the speechreading of sentences. Ear and Hearing. 1988; 9(6):306–312. [PubMed: 2975613]

Dorman M, Loizou P, Fitzke J, Tu Z. The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. Journal of the Acoustical Society of America. 1998; 104:3583–3585. [PubMed: 9857516]

Fabry D, Firszt JB, Gifford RH, Holden LK, Koch D. Evaluating speech perception benefit in adult cochlear implant recipients. Audiology Today. 2009; 21:36–43.

Gifford RH, Shallop JK, Peterson AM. Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. Audiology & Neuro-Otology. 2008; 13(3):193–205. [PubMed: 18212519]

Goldman-Eisler, F. Psycholinguistics: Experiments in spontaneous speech. New York: Academic Press; 1968.

Killion M, Niquette P, Revit L, Skinner M. Quick SIN and BKB-SIN, two new speech-in-noise tests permitting SNR-50 estimates in 1 to 2 min (A). Journal of the Acoustical Society of America. 2001; 109(5):2502-2502.

Nilsson M, Soli SD, Sullivan JA. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. The Journal of the Acoustical Society of America. 1994; 95(2):1085–1099. [PubMed: 8132902]

Peterson GE, Lehiste I. Revised CNC lists for auditory tests. Journal of Speech and Hearing Disorders. 1962; 62(27):62–70. [PubMed: 14485785]

Seneff, S.; Zue, V. Transcription and alignment of the TIMIT database. Proceedings of the Second Symposium on Advanced Man- Machine Interface Through Spoken Language; 20–22 November 1988; Oahu, HI. 1988.

Spahr AJ, Dorman MF. Performance of subjects fit with the advanced bionics CII and nucleus 3G cochlear implant devices. Archives of Otolaryngology--Head & Neck Surgery. 2004; 130(5):624–628. [PubMed: 15148187]

Spahr AJ, Dorman MF. Effects of minimum stimulation settings for the med el tempo+ speech processor on speech understanding. Ear and Hearing. 2005; 26(4 Suppl):2S–6S. [PubMed: 16082262]

Spahr AJ, Dorman MF, Loiselle LH. Performance of patients using different cochlear implant systems: Effects of input dynamic range. Ear and Hearing. 2007; 28(2):260–275. [PubMed: 17496675]

Thornton AR, Raffin MJ. Speech-discrimination scores modeled as a binomial variable. Journal of Speech and Hearing Research. 1978; 21(3):507–518. [PubMed: 713519]

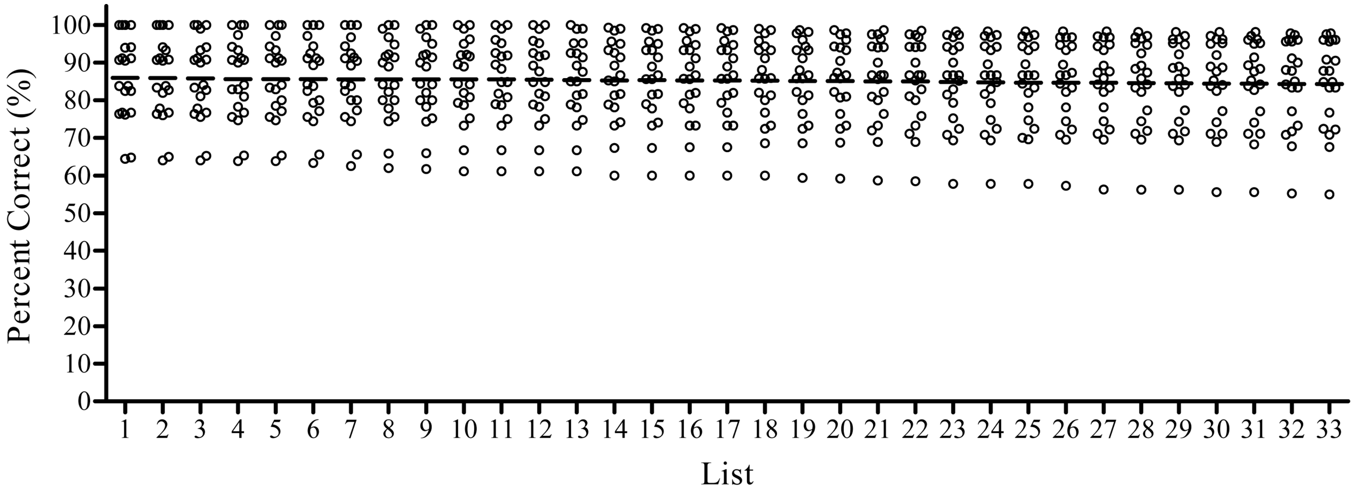

Figure 1.

Intelligibility estimates of the 33 sentence lists. Symbols represent the mean percent correct score of a single sentence presented to 15 normal-hearing subjects listening to a 5-channel cochlear implant simulation. The mean percent correct score for each list is indicated by a horizontal bar.

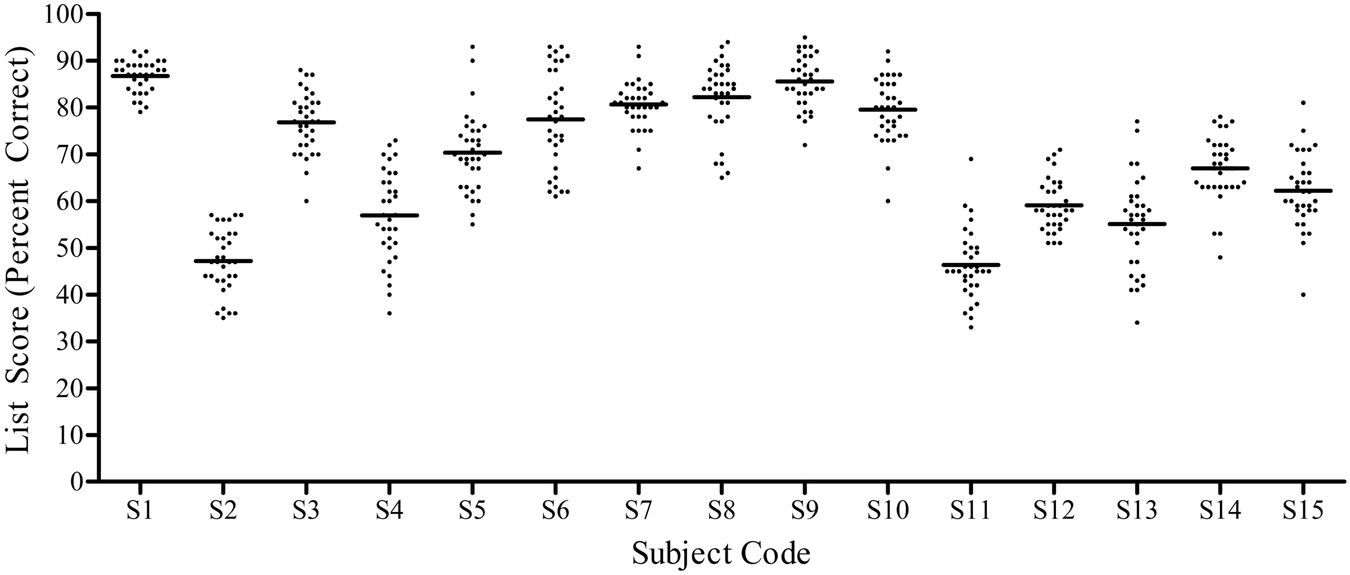

Figure 2.

List scores for 15 CI listeners. The absolute percent correct score for each of the 33 tested lists is shown as a closed circle. The mean level of performance for each listener is indicated by a horizontal line.

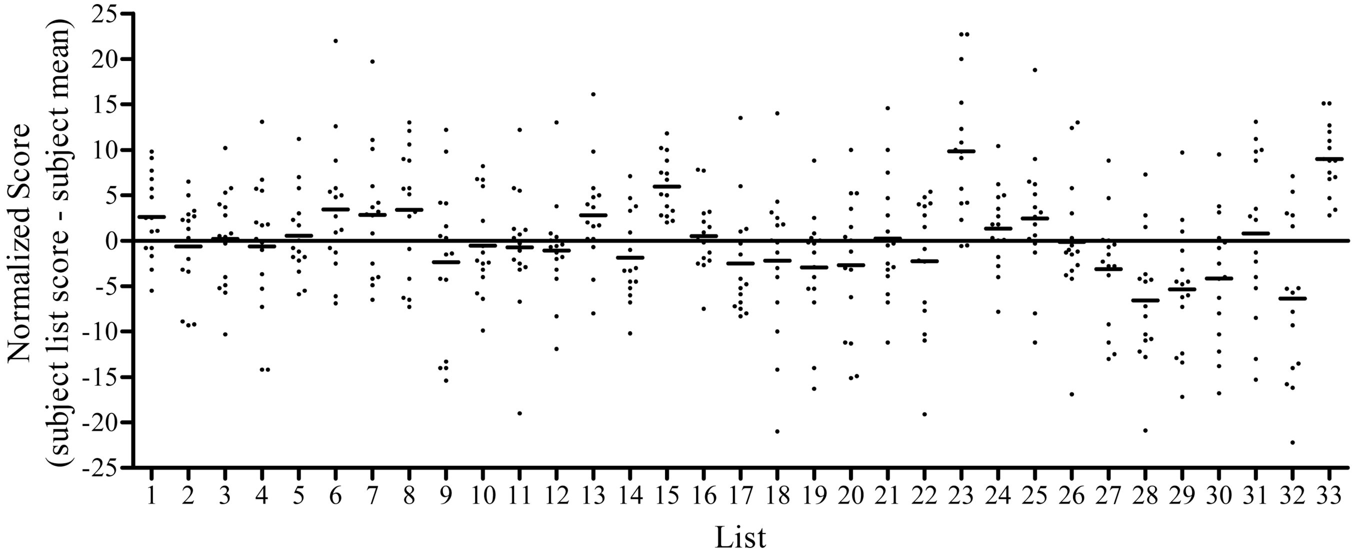

Figure 3.

Normalized scores for 15 CI listeners on all 33 sentence lists. Symbols represent an individual listener's list score relative to their overall mean level of performance. Positive values indicate better than average performance and negative values indicate below average performance. The average normalized score for each list is shown as a horizontal bar.

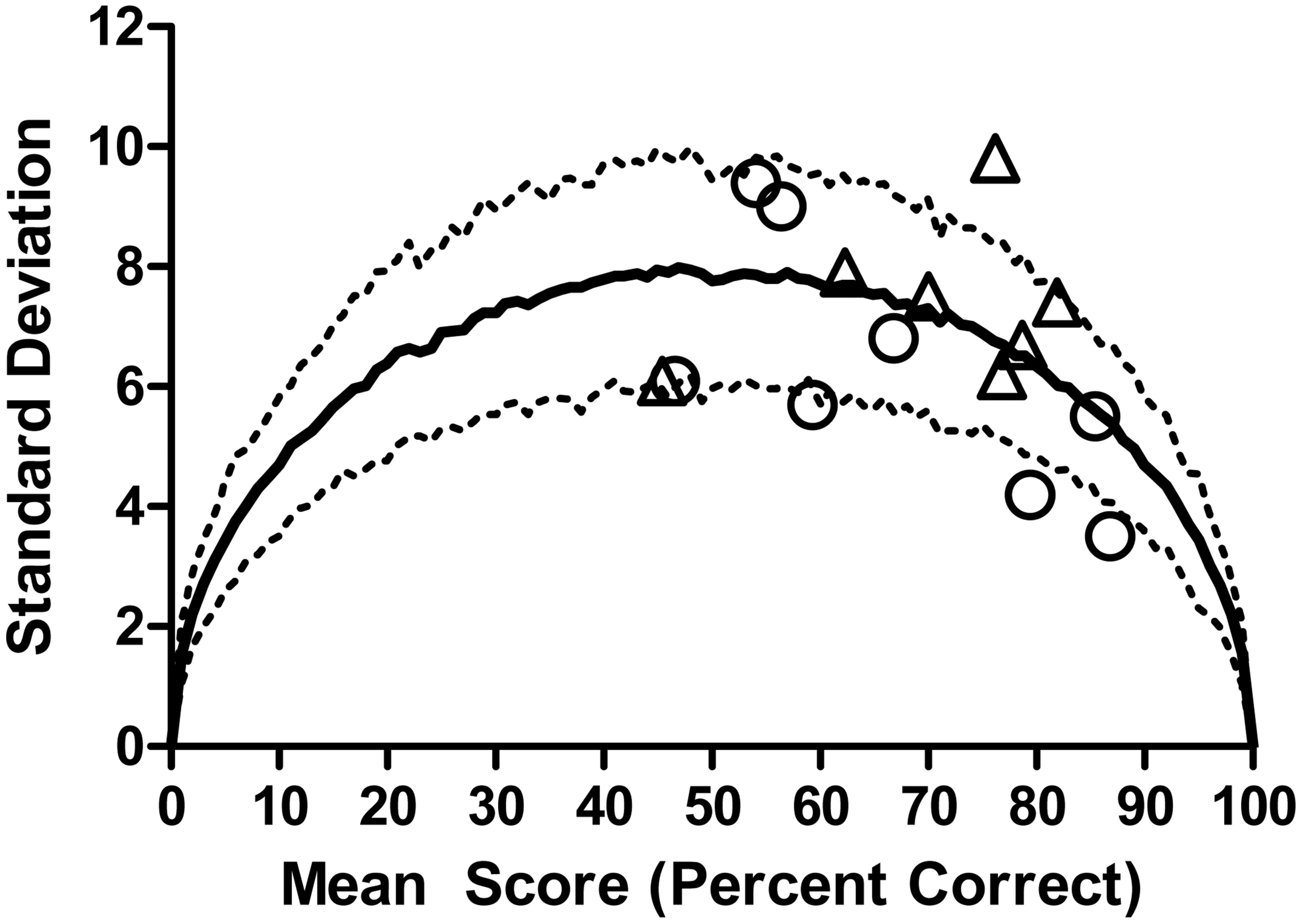

Figure 4.

Predicted variability of materials as a function of the mean percent correct scores. The predicted standard deviation (solid line) and 95% confidence intervals (dashed lines) are based on 29 list scores, with each list containing 40 items. Symbols represent the mean and standard deviation of scores from 15 cochlear implant listeners on 29 lists of AzBio sentences. For each subject, all lists were presented in quiet (circles) or in noise at a single signal-to-noise ratio (triangles) to prevent ceiling or floor effects.

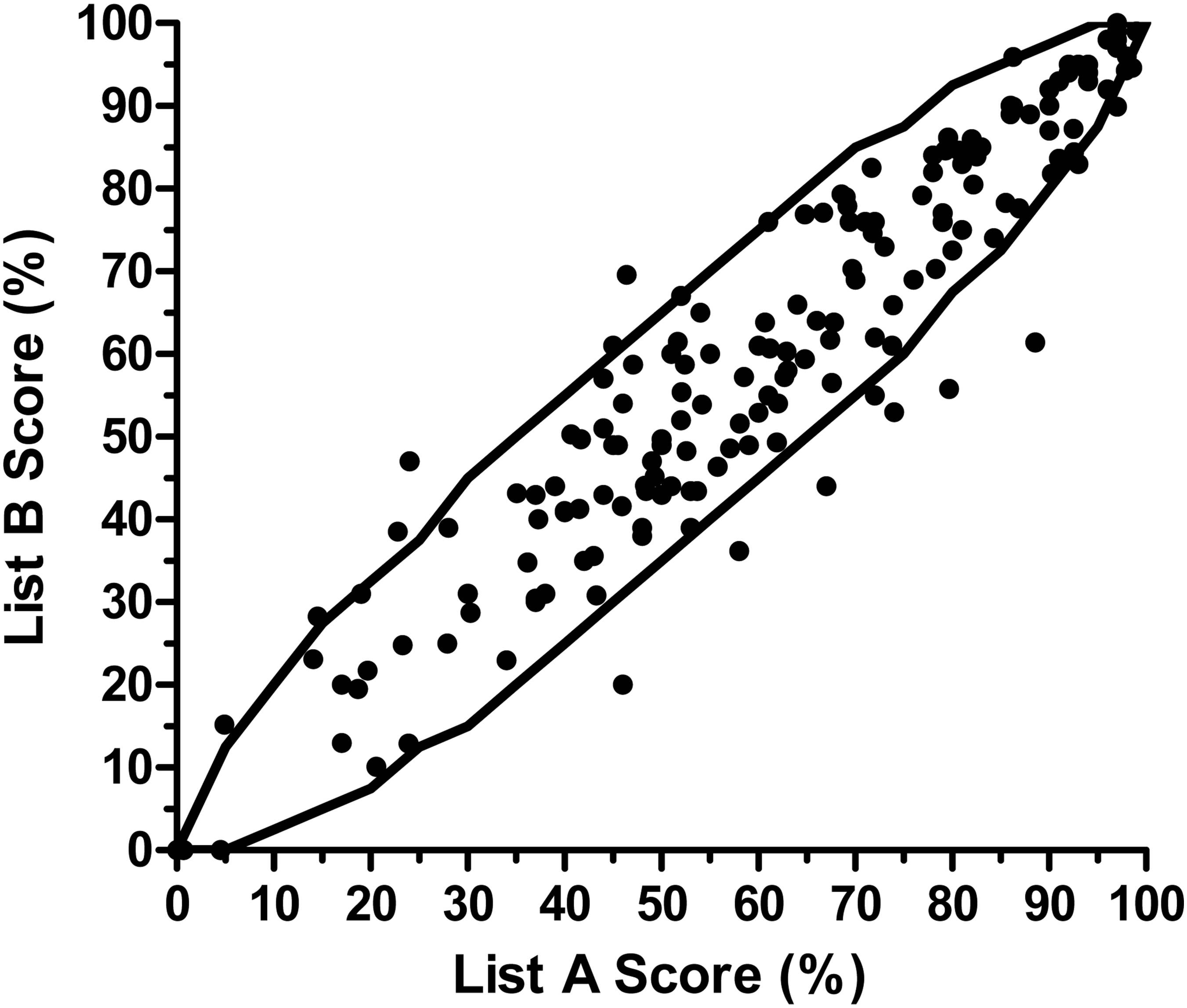

Figure 5.

Comparison of 172 instances where individual cochlear implant listeners (n=66) were tested on two lists within the same listening condition. Solid lines represent the upper and lower 95% confidence intervals for single list comparisons. Symbols represent scores on the first (A) and second (B) list tested within the same condition. Scores falling outside of the 95% confidence interval would be incorrectly labeled as significantly different.

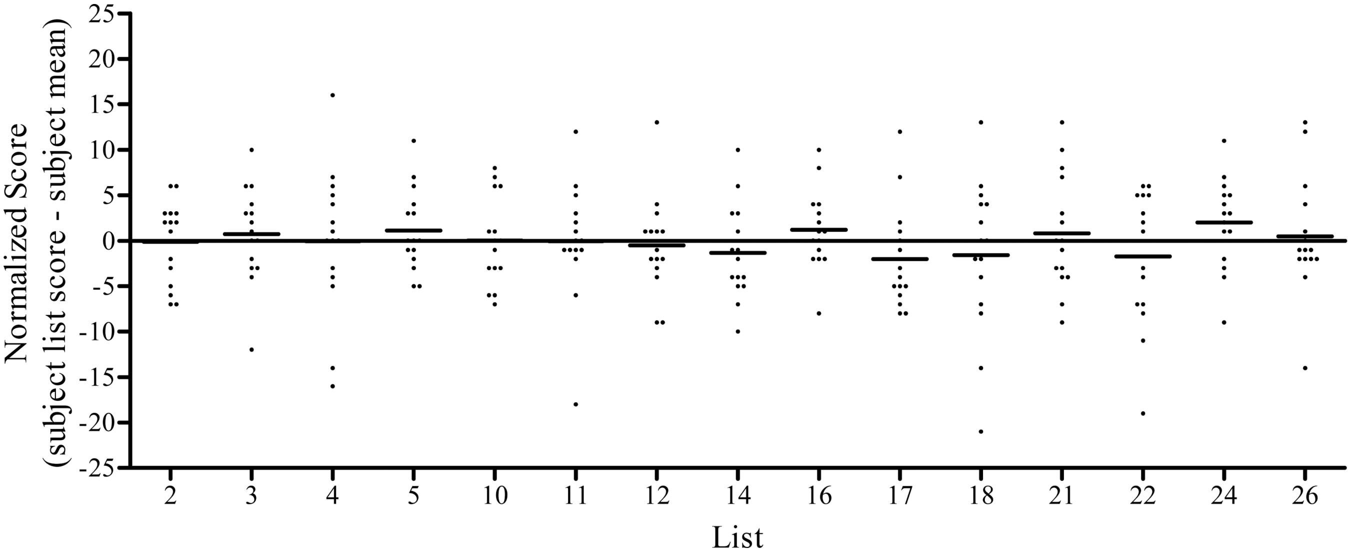

Figure 6.

Normalized scores for 15 CI listeners on the 15 sentence lists included in the AzBio Sentence Test. Symbols represent an individual listener's list score relative to their mean level of performance on all 15 lists. Positive values indicate better than average performance and negative values indicate below average performance. The average normalized score for each list is shown as a horizontal bar.

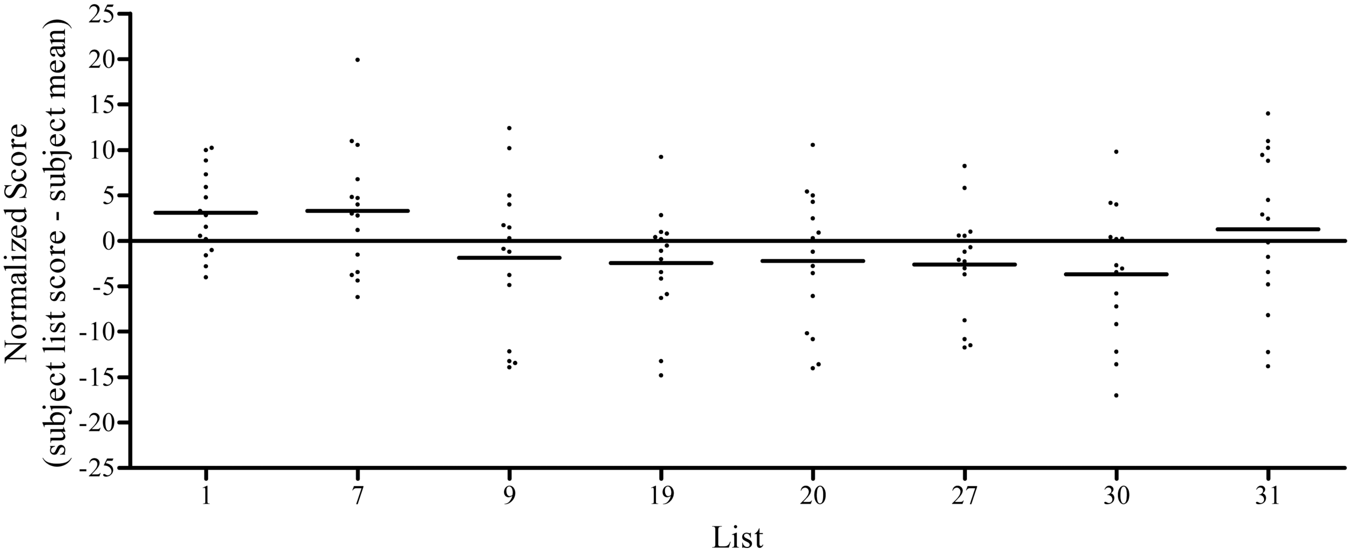

Figure 7.

Normalized scores for 15 CI listeners on the 8 sentence lists included in the Minimum Speech Test Battery. Symbols represent an individual listener's list score relative to their own mean level of performance on the 8 lists. Positive values indicate better than average performance and negative values indicate below average performance. The average normalized score for each list is shown as a horizontal bar.

Table 1

Upper and lower 95% confidence intervals for AzBio sentences lists computed using a binomial distribution model with 40-items per list. Lower and upper confidence intervals are shown as a function of starting level of performance (percent correct) when testing 1 or 2 lists per condition.

	1 list per condition		**2 lists per condition**	
Score	**Lower**	**Upper**	**Lower**	**Upper**
0	0	0	0	0
5	0	13	1	10
10	3	20	4	16
15	5	28	8	24
20	8	33	11	29
25	13	38	16	35
30	15	45	20	40
35	20	50	25	46
40	25	55	29	51
45	30	60	34	56
50	35	65	39	61
55	40	70	44	66
60	45	75	49	71
65	50	80	54	75
70	55	85	60	80
75	60	88	65	84
80	68	93	71	89
85	73	95	76	93
90	80	98	83	96
95	88	100	90	99
100	100	100	100	100