An algorithm that administer adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function

Gitte Keidser, Harvey Dillon, Jorge Mejia, and Cong-Van Nguyen.

National Acoustic Laboratories and the Hearing CRC

Corresponding author:

Gitte Keidser, NAL, 126 Greville St, Chatswood, NSW 2067, Australia;

Phone: +61 2 9412 6800;

Email: gitte.keidser@nal.gov.au

Keywords:

Algorithm, speech reception threshold, reliability, psychometric function, speech performance in noise

Abbreviations:

3FA = three-frequency-average; ANOVA = analysis of variance; BKB = Bamford-Kowal-Bench; HINT = hearing in noise test; HIP = hearing-impaired participant; HTL = hearing threshold level; ILTASS = International long-term average speech spectrum; NHP = normal-hearing participant; SE = standard error; SD = standard deviation; SNR = signal-to-noise ratio; SPL = sound pressure level; SRTn = speech reception threshold in noise

Abstract

Objective: To introduce and verify an algorithm designed to administer adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function.

Design: Speech-in-noise performances were measured using BKB sentences presented in diffuse babble-noise, using morphemic scoring. Target of the algorithm was a test-retest standard deviation of 1.13 dB within the presentation of 32 sentences. Normal-hearing participants completed repeated measures using manual administration targeting 50% correct, and the automated procedure targeting 25%, 50%, and 75% correct. Aided hearing-impaired participants completed testing with the automated procedure targeting 25%, 50%, and 75% correct, repeating measurements at the 50% point three times.

Study sample: Twelve normal-hearing and 63 hearing-impaired people who had English as first language.

Results: Relative to the manual procedure, the algorithm produced the same speech reception threshold in noise ($p = 0.96$) and lower test-retest reliability on normal-hearing listeners. Both groups obtained significantly different results at the three target points ($p < 0.04$) with observed reliability close to expected. Target accuracy was not reached within 32 sentences for 18% of measurements on hearing-impaired participants.

Conclusions: The reliability of the algorithm was verified. A second test is recommended if the target variability is not reached during the first measurement.

Introduction

Speech-in-noise testing is used clinically to assess a person’s hearing problem or to select the appropriate technology, and in research to evaluate different amplification schemes. To avoid floor or ceiling effects, an adaptive procedure for varying the speech or noise levels is commonly used to track the signal-to-noise ratio (SNR) that produces a target of 50% correctly recognised items (e.g. Levitt & Rabiner, 1967). This measure is referred to as the speech reception threshold in noise (SRTn). The adaptive procedure lends itself to automation, and several implementations for controlling the adaptive variation of speech or noise levels exist (e.g. Cheesman, 1992; Hagerman & Kinnefors, 1995; Brand & Kollmeier, 2002; Smits et al., 2004; Terband & Drullman, 2008; Ozimek et al., 2010; Dawson et al., 2011). Some of these implementations further handle the scoring of key items that are either selected from a list of alternatives, or directly entered by the study participant.

One advantage of automating the adaptive procedure is the possibility of handling more sophisticated decision processes within a shorter time frame to improve reliability of measurements. For example, a specified reliability of measurements could be ensured by monitoring the changes in SNR as the test progresses. If reliability is not adequate at the end of the administration of a regular list, the test could continue by automatically adding more sentences to the measurements. Contrary, if a person performs very reliably the test can be concluded earlier to save testing time. One potential disadvantage related to SRTn measurements is the chance of the SRTn being unrealistically poor and therefore for the result to have little real-life validity (Naylor, 2010). This problem can be addressed by aiming for a higher percentage point on the psychometric function, which would require a more complicated arithmetic to control the adaptive procedure that could be difficult to handle manually.

This report describes an algorithm developed to control the adaptive procedure for obtaining the SNR at selectable points within the range of 20 to 80% on the psychometric function whilst achieving desired test-retest reliability. Verification data obtained on normal-hearing and hearing-impaired listeners are presented.

The algorithm

The algorithm assumes speech material comprising of sentences where each sentence presents a set of key items (e.g. words, phonemes, or morphemes). It controls the adaptive changes to the level of either speech or noise depending on how many key items in each sentence are correctly identified; assuming the noise level is kept constant, the level of speech is increased if the ratio between correctly identified key items to the actual key items presented is less than a given target percentage point, decreased if the ratio exceeds target, and maintained if the ratio equals target. To improve reliability, the convergence of the algorithm to reach the SNR for a set target is based on three phases during which the step size gradually decreases according to a set of criteria:

In phase 1, iterations of 5 dB are used until at least 4 sentences are completed and at least one reversal is detected.

In phase 2, iterations of 2 dB are used until at least a further 4 sentences are completed and the standard error (SE), see equation 1, is no greater than 1.0 dB. The estimation of SE only includes the number of presentations completed in phase 2.

Equation (1) $SE \leq \frac{2 \times std(x)}{\sqrt{n}}$, where

$$std(x) = \left(\tfrac{1}{n-1}\right)\left(\sum_{i=1}^{n}\left(x_i - \widehat{x}\right)^2\right)^{\frac{1}{2}}$$

and

$$\widehat{x} = \left(\tfrac{1}{n}\right)\sum_{i=1}^{n} x_i$$

Where x is the presentation level or SNR, n is the total number of sentences tested, and i denotes each independent sentence.

In phase 3, iterations of 1 dB are used until at least a specified number of sentences (16 by default) are completed from the start of phase 2 and the SE is no greater than 0.8 dB. The estimation of SE includes the number of presentations completed in phases 2 and 3.

Equation 1 includes a correction factor of 2. SE is usually estimated from the standard deviation of a sample by dividing by the square root of the number of points in the sample. That calculation is based on the assumption that the points in the sample are independent of each other. When the points in question are the SNRs at which succeeding items are presented in an adaptive test, the points are not independent as each SNR can be different from the preceding SNR only by the selected step size. Monte Carlo simulation was used to estimate the degree to which the standard formula for SE under-estimates the actual SE under this condition. The Monte Carlo simulation consisted of 100 repetitions of a 30-run adaptive test with a fixed step size and a fixed underlying psychometric function. For each repetition, the estimated SE arising from a conventional calculation (i.e. equation 1 with the factor of 2.00 excluded) was calculated, and the squares of these values were averaged across the repetitions and the square root of this value taken. The true SE was also calculated based on the spread of the mean SRTn values across repetitions. Finally, the correction factor needed was calculated as the ratio of the true SE divided by the estimated SE. This process was repeated for different combinations of step size (1, 2 and 4 dB) and psychometric slope at the midpoint (5, 10 and 20 % per dB). Table I shows the correction factors calculated by this process when the 50% point on the function was targeted.

On theoretical grounds the degree of underestimation of SE should be greatest for small step sizes and shallow psychometric functions, which the Monte Carlo simulation confirmed. Also, on theoretical grounds, the degree of underestimation should depend only on the product of the step size and the psychometric function slope, which the simulation also confirmed. As the factor had to be chosen prior to the experiment being performed, as psychometric function slopes between 10 and 20% were expected, and as the SRTn calculation depended primarily on points collected with a 2 dB step size, a factor of 2.00 was chosen for inclusion within the test program. Further Monte Carlo simulations indicated that, as expected, somewhat higher correction factors would be appropriate at the 25 or 75% target points, because the local slope of the psychometric function is shallower there than at the 50% point. The factor was, however, left at 2.0 for all measurements, which means that greater test-retest variations than targeted can occur at the more extreme points on the psychometric function.

The stopping criterion of a SE less than 0.8 dB is an independent choice driven by the desired test-retest reliability. It should result in 95% of the scores of repeated tests falling within a 3.2 dB range. It can be made smaller with a consequence of increased test time, or the test time can be shortened if less accuracy is required. This implementation may require administration of more than one regular list of sentences to reach the target reliability, and the program does allow the user to select and specify the combination of lists used for a measurement and also includes a feature that will randomize the sentences within a regular list. An absolute maximum number of sentences can also be specified at which to conclude the test in case the SE appears not to reach the target. The implementation requires manual administration of the entry of number of correctly identified key items.

The platform

The program running the algorithm is developed to handle real time playback of speech and background noise through any combination of up to 52 loudspeakers. The software runs on a laptop or PC, with Microsoft Windows XP or above installed, that is connected to an ASIO supported (Steinberg Media Technologies GmbH, 2010) audio interface. Speech and noise are stored as multi-channel WAV files. To use the program, the experimenter specifies the audio configuration; i.e. map the speech and noise files to the appropriate channels of the interface, and hence loudspeakers in the test setup, and the parameters for the speech in noise measurements. The program also includes a calibration procedure to ensure the correct SPL from each loudspeaker in the test setup.

Verification

Participants: Twelve normal-hearing participants (NHP) who had English as their first language were recruited. The group comprised 7 females and 5 males who all had hearing threshold levels (HTL) better than 20 dB at the audiometric frequencies from 0.5 to 4.0 kHz, except one person who presented an HTL of 20 dB at 4.0 kHz in one ear.

Data were later obtained on 63 aided hearing-impaired participants (HIP) who took part in an unrelated study. This group comprised 24 females and 39 males who had symmetrical, sensorineural hearing loss. According to the three-frequency-average (3FA) HTLs, calculated across 0.5, 1.0, and 2.0 kHz, the degree of hearing loss varied from very mild (3FA = 25 dB HL) to moderately severe (3FA = 58 dB HL).

Test material and setup: Speech and noise consisted of the BKB sentences (Bench et al., 1993) and an eight-talker babble-noise extracted from the NAL CD of Speech and Noise for Hearing Aid Evaluation (Keidser et al., 2002). Both stimuli were filtered to match the International Long-term Average Speech Spectrum (ILTASS) by Byrne et al. (1994). The stimuli were presented from a computer through a multi-channel soundcard (Edirol FA-101) and two four-channel amplifiers (Yamaha XM4080) to five loudspeakers (Tannoy 800) positioned 1.5 metres from the centre of the array. Speech was presented from 0^0 azimuth. Noise was presented uncorrelated from $\pm 45^0$ and $\pm 135^0$ azimuths. For the NHPs the presentation level of the continuous background noise was kept constant at 62 dB SPL while the level of speech was changed adaptively, starting at 0 dB SNR. Directed by the original study aim, the frontal loudspeaker was moved to a distance of 1 m from the participants while the other loudspeakers were positioned 2 m from the centre of the array when testing the HIPs. Further, the noise level was fixed at 55 dB SPL and the starting SNR was 10 dB SNR.

The BKB lists are designed to yield 50 key words across 16 sentences. In this study, a morphemic scoring method was used in which every meaningful unit of each key word was considered a key item. For example, postman and boys each consist of two morphemes (‘post’ + ‘man’ and ‘boy’ + ‘s’). Morphemic scoring method (Bamford & Wilson, 1979) was chosen to increase the range of percentage points targeted on the psychometric function with the least number of sentences required. As shown in Figure 1 the average number of key items in each BKB sentence, when using morphemic scoring, is 6, resulting in an average of 96 key items in a traditional BKB list. As a result, percentage points ranging from 16% to 83% are feasible targets with the introduced adaptive procedure.

Protocol: The HIPs were fitted with a pair of either Siemens Motion 500 M, 501 SX, or 501 P. Gain was set to match the NAL-NL2 target (Keidser et al., 2011), the omnidirectional microphone mode was selected, and all adaptive features were disabled. Using the automated procedure, the SNRs needed to reach the 25%, 50%, (SRTn) and 75% points on the psychometric function were obtained. For NHPs measurements were completed twice, while the HIPs only repeated the SRTn measurement (three times). Two lists were combined for each of the automated measures allowing for a maximum of 32 sentences to be presented for each condition. The stopping criterion for the SE was 0.8 dB and the minimum number of sentences required before the automated procedure could be terminated was selected to be 13 for NHPs and 16 (default) for HIPs.

The SRTn was also measured twice using the conventional manual method on NHPs. One list (16 sentences) was administered for each of the manual measures which were preceded by one practice list to obtain an individual starting level of speech. During administration of the practice list, the initial step size of 5 dB was reduced to 1 dB after 3 reversals had been obtained. A step size of 1 dB was used for each of the actual test lists. Test conditions were presented in a balanced randomised order across participants.

Results: Tables II and III summarise the results for NHPs and HIPs, respectively. Firstly, where applicable, the mean SNR across repeated measures was calculated for each participant. From this data, the average SNR and standard deviation (SD) value across participants were obtained (i.e. the inter-participant SD). Secondly, the test-retest differences were calculated for each participant and the SD of the test-retest difference values obtained across participants (i.e. the intra-participant SD). Tables II and III show for each test condition, the average SNR, inter-participant SD, intra-participant SD, and N, where N is the average number of sentences needed to obtain the target SE when excluding the phase 1 sentences. For the HIPs, three intra-participant SDs were obtained from the three repeated measures, and the average of these three SDs is reported.

According to a repeated measures ANOVA using the measured SNRs as observation, and condition and repetition as repeated measures, normal-hearing data showed a significant effect of condition ($F_{3,33} = 22.8$, $p < 0.000001$). A Tukey HSD post hoc analysis revealed no significant difference between automated and manual SRTn measurements ($p = 0.96$), while the automated measurements at the three percentage points were all significantly different from each other ($p < 0.04$). Repetition was not significant ($F_{1,11} = 0.004$, $p = 0.95$) and neither did the interaction between condition and repetition reach significance ($F_{3,33} = 1.5$, $p = 0.24$). Both the inter-participant SD and the intra-participant SD were lower for the automated than for the manual measurements. The measured intra-participant SDs of 1.1 – 1.3 dB with the automated procedure agree well with that predicted from two measurements reaching a SE of about 0.8 dB (sqrt(2)*0.8 dB = 1.13 dB). Testing was terminated without the target SE being reached after presentation of 32 sentences in 4% of 72 measurements.

Using the hearing-impaired data as observations, a repeated measures ANOVA showed that the average SNRs measured for the three percentage points were significantly different ($p < 0.0000001$). There was a significant correlation between the measured SNR at each percentage point and 3FA hearing loss ($r > 0.57$; $p < 0.0001$) as people with higher degree of hearing loss required better SNRs at a rate of 2 dB for every 10 dB variation in hearing loss.

Consequently, greater inter-participant SDs were recorded for this group. The regression lines fitted to data at each percentage point are very close to parallel, suggesting that in the aided condition the slope of the psychometric functions do not change with degree of hearing loss. The average intra-participant SD measured at the SRTn was in good agreement with that measured for NHPs, although slightly higher. The percentage of measurements for which testing was terminated after presentation of 32 sentences, without reaching the target SE, was also higher at 18%, as can be inferred from Figure 2. This observation partly explains the higher intra-participant SD.

Discussion

Mean data collected on NHPs and HIPs showed that the algorithm produced valid data with variability in repeated measures that, independent of the target percentage point, approximated the expected value of 1.13 dB. The test-retest variability of group data slightly exceeded the expected value, especially for the HIPs. In the case of the HIPs, the shallower slope of the psychometric function around the SRTn partly explains the higher value. More generally, higher variability values likely occurred because for some individuals, the maximum number of 32 sentences was expended without reaching a SE below 0.8 dB. While this occurred in only 4% of all measurements with the NHPs, the incidence was close to 20% among the HIPs and involved 57% of the participants. In some of these cases, the algorithm had not even converged to phase 3 at the time 32 sentences had been presented. The behaviour was not consistent within certain participants and did not depend on degree of hearing loss, age, or specific lists. If disregarding the 24 HIPs for whom the target SE was not reached during at least one of the SRTn measurements, the intra-participant SD is reduced to 1.3 dB. Consequently, it is recommended to review the result after each measurement and to repeat the test, if the target reliability is not obtained (e.g. Dawson et al., 2011).

The intra-participant variations reported in this paper are directed by the selected SE, and was reached after presenting, on average, 17 sentences (ranging from 13 to 27 sentences) to NHPs and 20 sentences (ranging from 16 to 29 sentences) to HIPs. It should be noted that for a similar number of sentences (10-30), other studies have reported lower intra-participant SDs than obtained in this study (e.g. Hagerman, 1982; Nilsson et al., 1994; Hagerman & Kinnefors, 1995; Wagener & Brand, 2005). Further, the efficiency of our algorithm seems lower than expected. Using the definition proposed by Taylor and Creelman (1967), where efficiency is inversely proportional to the test-retest variability, and hence the number of sentences presented, efficiency was calculated to be 67% and 63% based on the data obtained for the NHPs and HIPs, respectively. These values are lower than the 80% estimated for a single threshold measurement, based on 30 sentences, when using an adaptive method aiming to improve convergence and stability through the use of variable step sizes (Brand & Kollmeier, 2002). There are two possible explanations for the discrepancies in intra-participant reliability and efficiency of the adaptive algorithm reported in this and other studies when a similar number of sentences is used. One is that in the past studies, testing was done over headphones, which eliminates variations due to shadowing effects resulting from head movements, which were possible during the free field testing in this study. Another is list equivalence (Dillon, 1982) that has not been refined in the version of BKB sentences used in this study. In contrast, efforts were made to equate sentence difficulty when developing the HINT sentences (Nilsson et al., 1994), and list equivalence is inherent in the matrix tests investigated by Hagerman (1982) and Wagener & Brand (2005), due to the limited number of words used.

Despite the fact that the same step size and a similar number of sentences were used to reach the SRTn in NHPs with the manual and automated procedures, a lower inter-participant variation was obtained with the automated procedure. This is likely due to elimination of some human errors that may occur when the administrator has to score the number of correct items, calculate the proportion of correct items, and change the presentation level within a short time frame.

Data obtained in this study on NHPs show a much steeper psychometric function (20%/dB SNR vs 10%/dB SNR) with a shifted SRTn of -5 dB than reported with the same material in Keidser et al. (2002). The discrepancy in slope is mostly a consequence of the method used in the two studies. Previously, the percentage correct items was measured for each individual at fixed SNRs, while in this study the SNR was obtained for each individual at a fixed percentage point on the psychometric function. Assuming that the slope of psychometric functions measured on normal-hearing listeners is similar, but that the position of the functions (SRTn) varies, then fitting a function to points extracted from the individually measured functions at fixed SNRs will display a shallower slope. This is because the upper and lower asymptotes are directed by the position of the psychometric functions obtained on the participants requiring the best and the worst SNRs, respectively, stretching the slope of the psychometric function between these points. If instead the function is fitted to data extracted at fixed percentage points on the individual functions, the common slope will be maintained, and the position of the fitted function will be in the middle of the range of the individually measured functions. The shift in SRTn is presumed to be primarily explained by differences in the environment under which the tests were conducted. Previously, speech and noise were presented co-located, whereas in the current study, speech and noise were spatially separated, and the noise was diffuse (e.g. Arbogast et al., 2002).

The information obtained in this study on the slope of the psychometric function could be used to fine tune the correction factor used to estimate the SE applicable to a single measurement. Specifically, the size of the correction factor (set to 2 for all measurements in this study) could be increased as the target point on the psychometric curve is moved away from the 50% point on the curve. It should further be noted that in a recent paper, Smits & Festen (2011) suggested that the optimal measurement procedure for testing speech in noise differs between normal-hearing and hearing-impaired listeners due to the psychometric function typically being shallower for hearing-impaired listeners. Thus, it is suggested that larger step sizes and a higher target point on the psychometric function should be used when testing hearing-impaired than when testing normal-hearing people. Such design issues, which would improve the efficiency values obtained with the present implementation, could be considered for future algorithms.

Conclusion

An algorithm was developed to control the adaptive procedure in speech-in-noise testing whilst achieving a target reliability at selectable points on the psychometric function. The algorithm was found to produce data similar to what is obtained with conventional manual administration using the same material, to reduce inter-participant variability in normal-hearing listeners, and to reach the target reliability in 96% and 82% of measurements on normal-hearing and hearing-impaired listeners. Improved reliability relative to manual administration was partly the result of automating the process and partly the result of building test reliability into the stopping criterion.

Acknowledgments

This work was financially supported by the Department of Health and Aging and the HEARing Cooperative Research Centre established and supported under the Cooperative Research Centres Program – an Australian Government initiative. We would like to thank Elizabeth Convery, Els Walravens, and Ingrid Yeend for helping out with data collection and the many volunteers who participated in the speech-in-noise testing.

References

Arbogast, T. L., Mason, C. R., Kidd, Jr., G. 2002. The effect of spatial separation on informational and energetic masking of speech. J Acoust Soc Am 112, 2086–2098.

Bench, R.J., Doyle, J., Daly, N., Lind.C. 1993. *The BKB/A Speechreading Test.* Victoria: La Trobe University.

Bamford, J., Wilson, I. 1979. Methodological considerations and practical aspects of the BKB sentence lists. Bench, Bamford (eds.) *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children*. London: Academic Press, pp 148–187.

Brand, T., Kollmeier, B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J Acoust Soc Am, 111(6): 2801-2810.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K. et al. 1994. An international comparison of long-term average speech spectra. J Acoust Soc Am, 96(4), 2108-2120.

Cheesman, M.F. 1992. An automated technique for estimating speech reception thresholds in multi-talker babble. J Speech-Language Pathol Audiol 16(3): 223-227.

Dawson, P., Mauger, S.J., Hersbach, A.A. 2011. Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients. Ear Hear 32(3): 382-390.

Dillon, H. 1982. A quantitative examination of the sources of speech discrimination test score variability. Ear Hear 3(2): 51-58.

Hagerman, B. 1982. Sentences for testing speech intelligibility in noise. Scand Audiol 11: 79-87.

Hagerman, B., Kinnefors, C. 1995. Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. Scand Audiol 24(1): 71-77.

Keidser, G., Dillon, H., Flax, M., Ching, T., Brewer, S. 2011. The NAL-NL2 prescription procedure. Audiology Research 1:e24: 88-90.

Keidser, G., Ching, T., Dillon, H., Agung, K., Brew, C., et al. 2002. The National Acoustic Laboratories' (NAL) CDs of Speech and Noise for Hearing Aid Evaluation: Normative Data and Potential Applications. Austr & NZ J Audiol 24(1):16-35.

Levitt, H., Rabiner, L.R. 1967. Use of a sequential strategy in intelligibility testing. J Acoust Soc Am 42: 609-612.

Naylor, G. 2010. Limitations of speech reception threshold as an outcome measure in modern hearing-aid research. Abstract from the International Hearing Aid Research Conference, Lake Tahoe: 19-20.

Nilsson, M., Soli, S.D., Sullivan, J.A. 1994. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am 95(2): 1085-1099.

Ozimek, E., Warzybok, A., Kutzner, D. 2010. Polish sentence matrix test for speech intelligibility measurement in noise. Int J Audiol 49(6): 444-454.

Smits, C., Kapteyn, T.S., Houtgast, T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol 43(1): 15-28.

Smits, C., Festen, J.M. 2011. The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: Steady-state noise. J Acoust Soc Am 130(5): 2987-2998.

Taylor, M.M., Creelman, C.D. 1967. PEST: Efficient Estimates on Probability Functions. J Acoust Soc Am 41(4): 782-787.

Terband, H., Drullman, R. 2008. Study of an automated procedure for a Dutch sentence test for the measurement of the speech reception threshold in noise. J Acoust Soc Am 124(5): 3225-3234.

Wagener, K.C., Brand, T. 2005. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurements procedure and masking parameters. Int J Audiol 44(3): 144-56.

Table I: The correction factors calculated by the Monte Carlo process when the 50% point on the psychometric function was targeted.

Step size (dB)	Slope of 5%/dB	Slope of 10%/dB	Slope of 20%/dB
1	4.9	3.5	2.3
2	3.3	2.3	1.6
4	2.3	1.6	1.2

Table II: The mean SNR, the inter-participant standard deviation, and the intra-participant standard deviation shown for the normal-hearing participants for four test conditions: Manual administration aiming for 50% identification, and automated administration aiming for 25%, 50%, and 75% identification. See text for definitions of the standard deviations. The last column shows the average number of scored sentences (N) needed across phases 2 and 3 in the automated administration to reach the target SE. A fixed number of 16 sentences was always used in the manual administration.

Condition	Mean SNR (dB)	Inter-participant SD (dB)	Intra-participant SD (dB)	N
Manual, 50%	-5.9	1.34	1.70	(16)
Automated, 25%	-7.7	0.94	1.13	19.6
Automated, 50%	-6.1	0.98	1.27	16.8
Automated, 75%	-5.2	1.19	1.27	15.5

Table III: The mean SNR and standard deviations (inter- and intra-participant) shown for the hearing-impaired participants for three test conditions: Automated administration aiming for 25%, 50%, and 75% identification. See text for definition of the standard deviations. The last column shows the average number of scored sentences (N) needed across phases 2 and 3 to reach the target SE.

Condition	Mean SNR (dB)	Inter-participant SD (dB)	Intra-participant SD (dB)	N
Automated, 25%	-0.7	3.06		21
Automated, 50%	1.0	2.85	1.41	20
Automated, 75%	2.5	3.61		21

Figure legends

Figure 1: The distribution of the number of morphemic key items in BKB sentences. The broken line shows the fitted normal function.

Figure 2: The distribution of standard errors of the signal-to-noise ratio obtained at the time of termination of the adaptive procedure for hearing-impaired listeners. Results are shown for three repeated measures (R1, R2, and R3) at the 50% point and for measures at the 25% and 75% points.

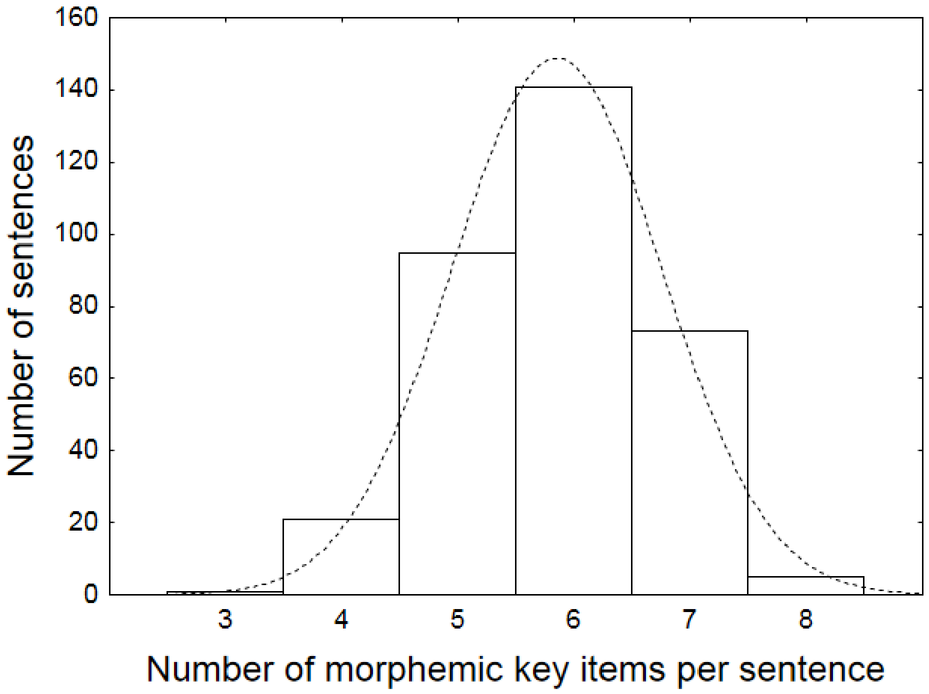
Number of sentences
0
20
40
60
80
100
120
140
160
3
4
5
6
7
8
Number of morphemic key items per sentence

Figure 1

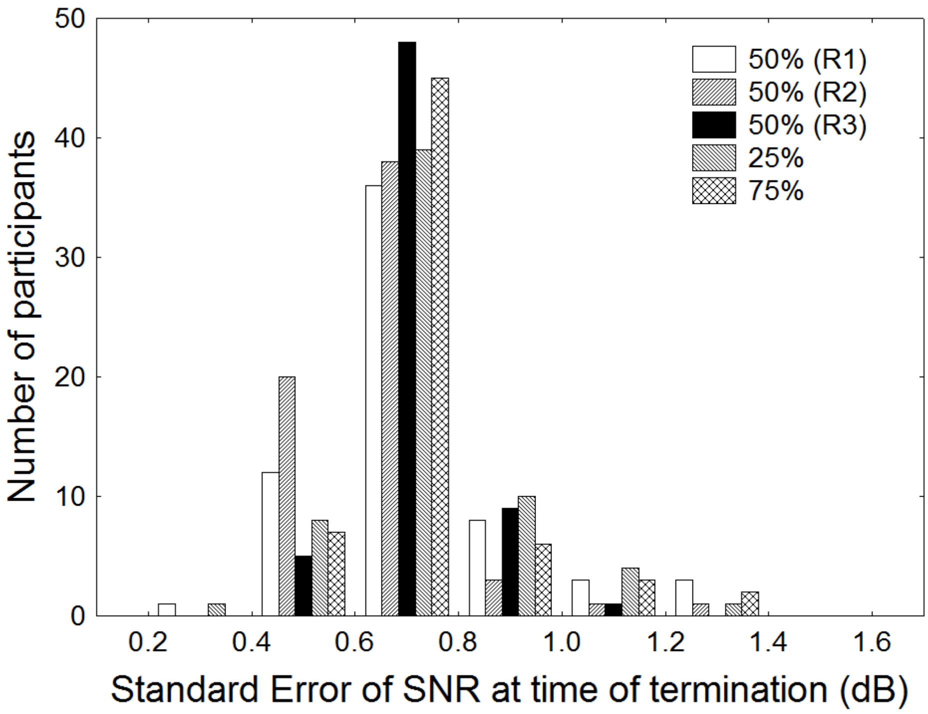
50% (R1)
50% (R2)
50% (R3)
25%
75%
Number of participants
0
10
20
30
40
50
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Standard Error of SNR at time of termination (dB)

Figure 2