TITLE

The development of an Afrikaans test for sentence recognition thresholds in noise

AUTHORS

Marianne Theunissen[a,b]
Johan J Hanekom[b]
DeWet Swanepoel[a,c]

AFFILIATION

[a] Department of Communication Pathology, University of Pretoria
[b] Department of Electrical, Electronic and Computer Engineering, University of Pretoria
[c] Callier Center for Communication Disorders, University of Texas at Dallas, USA

KEY WORDS

Sentence recognition
Speech reception / recognition
Speech audiometry
Speech-in-noise
Speech discrimination in noise
South Africa
Afrikaans

ABBREVIATIONS

HINT:	Hearing In Noise Test
BKB:	Bamford-Kowal-Bench sentences
SNR:	Signal-to-noise ratio
SNR-50:	Signal-to-noise ratio where 50% correct speech recognition is attained

CORRESPONDING AUTHOR

Prof D Swanepoel
Department of Communication Pathology
University of Pretoria
Lynnwood Road
Pretoria
South Africa
0002

E-mail: dewet.swanepoel@up.ac.za

ABSTRACT

Objective: The development of a valid and reliable Afrikaans test of sentence recognition thresholds in noise.

Design: A collection of sentences was developed, rated for naturalness and grammatical complexity, and digitally recorded using a female speaker. Sentences found to have similar psychometric curve slopes, with equivalent intelligibility at three different noise levels, were arranged into 22 phonemically matched lists of ten sentences each. List equivalence was evaluated in normal-hearing listeners in full and reduced bandwidth conditions. Test-retest reliability of the remaining lists was evaluated in a second group of listeners.

Study sample: All listeners were native speakers of Afrikaans with normal hearing. For evaluation of list equivalence, ten listeners were used. Twenty other listeners were used to evaluate test-retest reliability.

Results: A collection of eighteen phonemically matched lists was produced. Lists were found to be of equivalent difficulty in full and reduced bandwidth conditions, and to have good test-retest reliability in normal-hearing listeners. The average recognition threshold of these lists was -2.73 dB signal-to-noise ratio (standard deviation = 0.64 dB), and within-subject variability was 1.22 dB.

Conclusions: The developed test provides a valid and reliable means of measuring sentence recognition thresholds in noise in Afrikaans.

INTRODUCTION

The most common complaint of patients with sensorineural hearing loss is difficulty in understanding speech in situations with background noise (Smits et al, 2006). Although the basic audiometric test battery includes measures of speech reception or recognition, tests that could measure or predict a patient's ability to understand speech in noise are not routinely conducted (Killion & Niquette, 2000). However, because many patients have great difficulty in recognizing speech in the presence of

noise, the direct assessment of this skill could contribute significantly to effective counselling and selection of amplification options (Killion et al, 2004).

Tests of speech recognition using sentence material as stimulus tend to give a more global impression of an individual's speech perception than single words (Lutman, 1997), and a number of recent studies have therefore focused on the development of a test for speech recognition in noise using sentence materials. A prominent study in this regard was the development of the HINT (Hearing In Noise Test) by Nilsson et al (1994). Many researchers have since developed similar tests in other languages, such as German (Kollmeier & Wesselkamp, 1997), Dutch (Van Wieringen & Wouters, 2008; Versfeld et al, 2000), and French (Luts et al, 2008) or adapted the original HINT in different languages, including Brazilian Portuguese, Brazilian, Canadian French, Cantonese and eight other languages (Soli & Wong, 2008).

In South Africa, audiologists face the challenge of delivering services to a cultural and linguistically diverse patient load. There is a shortage of standardised speech audiometry materials and procedures in most of the eleven official languages, which prevents optimal audiological assessment. This shortage of test materials is particularly evident for sentence materials. In view of this dearth of materials for speech audiometry in South Africa, a project was undertaken to develop a test of sentence recognition in noise in one of the official languages. The language selected for the current project was Afrikaans, the third most common home language in the country, spoken by 13.3% of the population (Statistics South Africa, 2001) and the most common home language in the Gauteng province where the research was conducted (Central Statistical Service, 1995). Since there are no standardised or published tests of sentence recognition in noise in Afrikaans, the development of such a test could provide a valuable tool for the assessment of Afrikaans-speaking

individuals with peripheral hearing loss and different amplification devices, as well as listeners with auditory processing disorders.

METHOD AND RESULTS

The aim of the study was to develop a valid and reliable Afrikaans test for sentence recognition thresholds in noise. The procedures followed in the development of the test were developed after thorough examination of existing literature on the different variables involved in the test (Theunissen et al, 2009). The development of the American English HINT (Nilsson et al, 1994) and subsequent adaptations of the HINT in other languages were particularly valuable in guiding the planning of the research. The research process consisted of five distinct phases, each with its own aim, participants and procedures. The procedures and results for each of the distinct phases are described in this section. All the subjects used during this study were native speakers of Afrikaans. Participation in the study was voluntary and subjects were not paid for participation. Ethical clearance was obtained from the relevant ethics committee at the university prior to commencement of experimentation.

Phase I: Development of sentence material

Procedures

The aim of Phase I was to develop a suitable collection of recorded Afrikaans sentences for the assessment of speech recognition in noise. An examination of the literature showed that no formally standardised sentence tests or suitable sentence collections were available in Afrikaans. Therefore, a large collection of sentences was compiled using two methods, namely translation of existing material and compilation of original material similar in content and structure to the translated material.

The material selected for translation consisted of a large set (n = 336) of short sentences that was designed for use with British children, the BKB or Bamford-Kowal-Bench sentences (Bench & Bamford, 1979). These sentences were selected because of the size of the collection as well as their syntactic and grammatical simplicity. Additional sentence material similar in structure and content was developed using vocabulary known to be comprehensible to young children as a basis for the creation of new sentences (Vaillancourt et al, 2005). This method was used to ensure that the keywords in the sentences were of equal, known difficulty. Two collections of words considered to be suitable for evaluating young children were chosen for the purpose of expanding the sentence collection. The first collection was the “Afrikaanse Reseptiewe Woordeskattoets” (Afrikaans Receptive Vocabulary Test) (Buitendag, 1994). The second source of vocabulary was the “Phonetically Balanced Word Lists for Children” (Tesner & Laubscher, unpublished). Although these lists have not been formally standardised and remain unpublished (Tesner, personal communication, September 4, 2006), they have been successfully used for the evaluation of word recognition in young children for the past 40 years and were therefore considered a suitable resource for the compilation of sentence material.

After compilation, the sentence material was submitted to a group of native speakers of Afrikaans that represented different age groups, as well as a wide variety of educational and geographical backgrounds, for evaluation of naturalness. Naturalness was quantified by asking participants to rate sentences on a seven-point scale (1 = artificial; 7 = natural). This was done to ensure that the developed material was considered acceptable and natural by the general population (Nilsson et al, 1994; Wong & Soli, 2000; Vaillancourt et al, 2005; Hällgren et al, 2006). The total collection of 518 sentences was first submitted to five participants. Sentences that received a rating below 6 from at least two candidates were reviewed and edited according to suggestions provided by participants. Modified sentences were then

submitted to a second group of native speakers (n = 5) and sentences that still received a rating below 6 (n = 3) were excluded from the collection.

In order to determine the complexity of the material, the remaining sentences were submitted to an expert in language development and analysis with a PhD in Communication Pathology, for rating of grammatical complexity. The complexity of the grammatical structure of each sentence was rated on a seven-point scale, according to the age level of the clauses and phrases in the sentence. Subsequently, a panel of judges consisting of two speech therapists and two audiologists selected a suitable speaker for the recording of the sentence material. This speaker was a 26-year old female speech and language therapist judged to have good articulation, intelligibility, voice quality and resonance, as well as an appropriate speech rate and intonation, without a specific accent or over-articulated, unnatural speech quality.

The sentence material was recorded digitally in a sound-proof booth, using a Creative Labs Soundblaster Extigy external sound card (sampled at 44.1 kHz with 24-bit resolution). A Sennheiser ME62 microphone was placed on a microphone stand, 20 cm from the speaker's mouth. Each of the 515 sentences was recorded and saved as a separate .wav file. The speaker aimed to articulate all words clearly, without distortion of any sounds, and attempted to place equal emphasis on all parts of the sentence and maintain vocal effort throughout each sentence (Versfeld et al, 2000), while still retaining a natural intonation pattern. Because there are many varieties (standard and non-standard) of Afrikaans (Carstens, 2003), the speaker in the current recordings aimed for a standard pronunciation by avoiding the abnormal and striving for the general form as consistently as possible (Le Roux and Pienaar, 1976). After completion of the recordings, waveforms were edited using Praat software (Boersma & Weenink, 2006). Unwanted silences preceding and following the recorded speech were eliminated and the intensity of each sentence was re-

scaled to 70 dB before saving it to hard disc in .wav format. The average speaking rate of the sentences was 3.4 syllables, or 2.6 words, per second.

Results

In total 518 sentences were compiled, the majority of which (65%; n = 336) were translations of the BKB sentences; the remaining 35% were compiled from the vocabulary sources described earlier. These sentences had an average length of 5.6 words or 7.1 syllables. The minimum number of syllables and words was four, the maximum number of words was eight, and maximum number of syllables was nine. Five native speakers rated these sentences for naturalness on a scale of 1 to 7 and provided suggestions for change where considered necessary. One sentence was rejected, as the changes suggested by different participants were irreconcilable. Twelve other sentences received a rating lower than 6 from more than one candidate and were altered according to their suggestions. These twelve sentences were submitted for a second round of rating. During this round, only two sentences received an average rating lower than 6 and were therefore excluded from the collection. After two rounds of naturalness rating, 515 sentences remained. The grammatical complexity of the material was also rated, and the majority of the sentences (82%) received a rating between 1 and 3 on the seven-point scale. These three levels all indicate an age level of three, with level one indicating that all phrases and clauses in the sentence occur at or before three years of age, and level two and three indicating the presence of one or two phrase structures that occur after the age of three. Only 18% of the total collection received a grammar rating of between 4 and 7 (age level four to five, with none to two additional complex phrases).

Phase II: Selecting an equivalent subset of sentences

Method

The aim of this phase was to select from the total collection of recorded sentences those that are equally intelligible in the presence of noise. This implied selecting sentences that yielded both similar performance at specific signal-to-noise ratio (SNR) conditions and similar psychometric slopes (percentage intelligibility as a function of SNR). The total research sample for Phase II consisted of 22 individuals (11 male, 11 female), with an average age of 24 years, and ages ranging from 18 to 29 years. Ten subjects (five male, five female) participated in the first equalisation procedure, while twelve participants (six male, six female) took part in the second procedure. All participants were required to have normal hearing (hearing thresholds ≤ 15 dB HL at 250, 500, 1000, 2000, 4000 and 8000 Hz; normal otoscopic results; Type A tympanogram) and a negative otologic history. To control for the possibility of auditory processing and/or language disorders, subjects were required to have a minimum academic qualification of Grade 12 in a mainstream Afrikaans school, and no history of neurologic disease or injury, neurosurgery, childhood auditory processing disorder, or self-reported severe difficulty to hear speech in noise (Bellis, 2003). In order to prevent any effects that age might have on hearing and/or auditory processing, selection criteria stipulated that all subjects had to be between 18 and 30 years of age.

Speech-weighted noise with a spectral envelope matching the average power spectral density of the entire set of recorded sentences (Nilsson et al, 1994; Hällgren et al, 2006) was generated in a commercial software package for mathematics, and added to the recorded speech material. The sentences with the added noise were each saved as a separate wave file. Subsequently, two procedures aimed at the equalisation of the sentence collection were followed.

The first equalisation procedure was aimed at reducing the initial number of sentences and increasing the equivalence of the remaining material by eliminating those sentences that were significantly easier or harder than the majority. This was achieved by presenting the entire collection of 515 recorded sentences to ten different subjects in the presence of a fixed amount of background noise. For this procedure, the noise was added to each sentence so as to ensure an SNR of –5 dB (noise 5 dB louder than speech). This level was chosen in accordance with the findings of previous researchers (Plomp & Mimpen, 1979; Nilsson et al, 1994; Kollmeier & Wesselkamp, 1996; Versfeld et al, 2000; Wong & Soli, 2005; Vaillancourt et al, 2005; Hällgren et al, 2006; Wong et al, 2007), who indicated that an SNR of –5 dB should yield an intelligibility score (defined in this instance as the percentage of syllables correctly discerned) of approximately 50%. Scoring each syllable provided a means of determining a percentage intelligibility score for each separate sentence. This is similar to the word-scoring method used by previous researchers during the early phases of test development (Nilsson et al, 1994; Vaillancourt et al, 2005; Hällgren et al, 2006). Syllables rather than words were scored because the spelling rules of Afrikaans dictate that conjunctions are written as one word (the so-called conjunctive method) as opposed to the tendency to write conjunctions as two words (the disjunctive method) in English (Carstens, 2003). For this reason, there may be many multi-syllabic conjunctions that could receive more precise scoring if syllable scoring were used. A 50% target score was selected in an attempt to avoid the ceiling or floor effects that a higher or lower score might have on the procedure, especially since the difficulty level of the sentences was largely unknown at this stage.

Each subject was tested separately and was seated in a sound-proof booth with the test administrator (a qualified audiologist). All the sentences were presented to each subject at an SNR of –5 dB. The order of presentation of the sentences was

counterbalanced by arranging sentences into 10 playlists before testing, and starting each subject on a different playlist. The .wav files were presented using Praat software (Boersma & Weenink, 2006) as an interface, and the signal was routed to the auxiliary input of a Madsen Midimate 622 diagnostic audiometer via a Creative Labs Soundblaster Extigy external sound card (sampled at 44.1 kHz with 24-bit resolution). The sound was presented binaurally, using standard TDH39 headphones as transducer, at an intensity of 70 dB SPL. The sentences were presented one by one (controlled by the test administrator), and frequent breaks were given in between to prevent exhaustion on the part of the subject from affecting results. Each sentence was presented once only. Subjects were instructed to repeat what they had heard every time, even if it was only part of a word or sentence, and were encouraged to guess at the content if uncertain. Each response was compared to a text version of the sentence and the number of syllables repeated correctly was recorded on the test form.

Using these results, a mean percentage of intelligibility at SNR-5 (abbreviation of SNR at -5 dB) could be calculated for each sentence individually, as well as an overall mean of all the sentences, based on the percentage of syllables repeated correctly. A sentence was selected for the second equalisation procedure if its average score across participants fell within one standard deviation of the overall mean. The same equipment and procedures were used during the second equalisation procedure, with the exception of the SNR at which the material was presented. The first six subjects (three males and three females) listened to the 330 sentences selected during the first equalisation procedure at an SNR of –8 dB (noise 8 dB louder than speech), while the last six subjects (three male, three female) listened to the recordings at an SNR of –2 dB. A list of practice sentences was presented to each subject before commencing the test. Sentences were arranged into six playlists. Every subject listened to all six playlists, and the order of

presentation of the different playlists was counterbalanced so that each subject heard a different list first.

The -8 and -2 dB SNRs were selected to provide an indication of the accuracy of speech recognition at SNRs 3 dB worse and 3 dB better than the –5 dB SNR, as this could provide an approximation of the psychometric curve of each sentence, from which the slopes of these psychometric curves could be estimated. The three data points (percentage of syllables repeated correctly) at -8 dB, -5 dB and -2 dB generally covered the extent of the psychometric curve and were adequate to provide reliable estimates of the psychometric curve slopes (Kollmeier & Wesselkamp, 1997). The three data points for each sentence were fitted to an s-shaped approximation to the psychometric curve, the Gaussian cumulative distribution function that was also used by Versfeld et al (2000),

$$\Phi(r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{r-\mu}{\sigma}} e^{-t^2/2} dt , \qquad (1)$$

where r is the SNR and Φ is the percentage correct. The parameters μ and σ completely determine the shape of this s-shaped approximation to the psychometric curve. The 50% correct point on the curve corresponds to the parameter μ, permitting estimation of the 50% recognition threshold for each sentence (SNR-50). Parameter σ characterizes the spread of data at the SNR-50 point. The slope S at the 50% point of the psychometric curve is related to parameter σ via the expression

$$S = \frac{100}{\sigma\sqrt{2\pi}}. \qquad (2)$$

Slope S was estimated for each sentence with a curve-fitting algorithm that minimized the mean square error between the available data points and the approximation to the psychometric curve, eq. 1. A subset of sentences with similar

psychometric slopes and similar results at specific SNRs was then selected. Sentences with slopes that fell within one standard deviation of the mean slope were used in subsequent phases.

Note that the majority of previous reports on development of similar tests in other languages (see Soli & Wong, 2008 for a list of these reports) calculated or estimated this performance intensity slope value in a similar manner (using three data points), the objective being to guide RMS intensity adjustments required to equalize the difficulty of all the sentences in the collection. In the current project, however, these slopes were used to select a number of sentences with similar SNR at the 50% intelligibility point and similar slopes (i.e. similar intelligibility at other SNRs as well) for further testing. The only point of importance for the application of the test in its final format (i.e., finding the SNR-50) is the 50% point. However, the location of this 50% point was unknown at this stage of the test development, as in the final test format testing was conducted using an adaptive procedure with whole sentence (1 or 0) scoring instead of scoring the percentage of syllables correctly identified. It was therefore necessary to ensure the uniformity of the selection of sentences not only at the SNR-50 point, but across the extent of the psychometric curve, so that sentence lists could be compiled from a relatively uniform collection of sentences.

Results

The average percentage of syllables discerned correctly at SNR-5 by all participants for all sentences was 52%. The standard deviation from this mean was 27%. 330 sentences fell within one standard deviation of this mean, and were subsequently presented to a second group of participants at SNR-8 and SNR-2. The mean percentage of intelligible syllables for all sentences across all subjects at SNR-8 was found to be 18.2%, with a standard deviation of 12.7%. At SNR-2, the average percentage was found to be 89.5%, with a standard deviation of 11.0%. The slopes

of these 330 sentences were then estimated and ranged from 5.49% / dB to 52.13% / dB, with an average of 16.87% / dB and a standard deviation of 7.37% / dB. A subset of 220 sentences with slopes that were within one standard deviation from the mean slope was selected, having a mean slope of 16.04% / dB and standard deviation of 3.33% / dB. These 220 sentences that fell within the stipulated criteria were used in Phases III and IV of the project.

Phase III: Compilation and evaluation of lists

Method

Phase III was aimed at the development and evaluation of equivalent sentence lists. Speech material used in the determination of threshold levels should be different for each trial since understanding becomes easier if material is re-used or repeated (Nilsson et al, 1994). By implication, a test using sentence material should have a collection of different sentence lists that are of equal difficulty (equivalence). This was accomplished in the present study with the use of the well-reported method of arranging sentences into phonemically equivalent lists (Plomp & Mimpen, 1979; Nilsson et al, 1994; Hällgren et al, 2006; Vaillancourt et al, 2005, Wong & Soli, 2005).

The sentences selected through experimentation and statistical analyses in Phase II were submitted to an expert in speech analysis for phonetic transcription. Subsequently, the total number of occurrences of each phoneme was determined and then divided by the total number of lists to be compiled (22). This was then considered the target occurrence of this phoneme in each list. Any deviance from this ideal number was considered an error. A process of substitution was conducted to compile 22 lists with the lowest possible number of errors on all the phonemes. The total number of phoneme counts for all the lists was 1 078 (22 lists x 49 phonemes).

The errors on each of these counts were calculated as the difference between the ideal occurrence and the actual occurrence.

Following list compilation, the material was subjected to experimentation to evaluate its reliability or inter-list equivalence. Subjects in this phase adhered to the same selection criteria stipulated for participants in Phase II. The 10 subjects (five male, five female) had an average age of 24 years (ranging from 19 to 28 years). The same equipment and test environment described for Phase II were used in Phase III.

Scoring each syllable, as was done in earlier phases, provided a means for determining a percentage intelligibility score for each separate sentence. These percentages provided the data points necessary for the estimation of the slopes of the psychometric curve of each sentence. However, organizing sentences into lists to obtain similar difficulty across lists allowed the use of whole sentence scoring in phases III, IV, and IV. This method of scoring does not provide an intelligibility score for each sentence, but provides a more efficient method to determine an intelligibility score for a list of sentences. Test administrators could make a quick decision about the correctness of the sentence (1 or 0) and adjust the presentation level of the following sentence in the list accordingly.

The adaptive up-down presentation method (Plomp & Mimpen, 1979) was followed, where the first sentence in a list was presented at an SNR that would result in recognition below 50% (-8 dB in this case). This sentence was repeatedly presented at SNR levels that were increased (improved) in 2 dB-steps until the subject was able to repeat the entire sentence correctly. The test administrator compared the listener's repetition with a text version of the sentence, and all the words in a sentence had to be repeated accurately in order for it to be considered correct. Once the first sentence was correctly repeated, the next sentence was presented at the same

SNR. The presentation levels of subsequent sentences were determined each time by the correctness of the preceding sentence's repetition. If a sentence was repeated correctly, the following sentence was presented at a more difficult SNR (speech level decreased by 2 dB with noise level kept constant at 70 dB SPL). If a sentence was repeated incorrectly, the following sentence was presented at a better or easier SNR (speech level increased by 2 dB).

Testing was preceded by the presentation of two practice lists, each consisting of 10 randomly selected sentences that were rejected in the final experiment of Phase II. The order of list presentation was counterbalanced between subjects, with the first subject starting with list number 1, the second with list number 4, and all subsequent subjects starting with the next even-numbered list. At the end of a list of 10 sentences, the software calculated the SNR-50 (threshold level of SNR where 50% intelligibility would be attained) as an average of the presentation levels of the fifth to eleventh sentences (although an eleventh sentence was never presented, its presentation level could be determined according to the correctness of the tenth sentence's repetition). Although thresholds determined in this way could differ from those determined using syllable scoring, this does not affect the ultimate reliability of the test material, as the inter-list reliability was assessed and validated using this method during Phases III, IV and V.

Results

The goal of phonemic matching between lists was to attain an optimal arrangement of the sentences so that each list had the smallest number of errors in phonetic balance without jeopardising the balance of another list. The total number of errors for each list was calculated as the number of times that a phoneme deviated one or more from the ideal count. The average number of errors per list was 11.8, with a minimum of eight, and a maximum of 18 errors per list. The total number of phoneme

counts for all the lists was 1078 (22 lists x 49 phonemes). The errors on each of these counts were calculated as the difference between the ideal occurrence and the actual occurrence of a phoneme. 379 (35%) of these counts showed an error value of 0, that is, these phoneme counts were exactly equal to their ideal occurrence. Of the total number of errors for all phonemes across all lists, 83.2% fell within +/- 1 phoneme of its ideal occurrence. The maximum error for a single phoneme was + 5 (shortage of 5), but this occurred only once (less than 0.1% of total phoneme counts). The mean SNR-50 (sentence recognition threshold) of these 22 lists measured across subjects (n = 10) using the adaptive procedure and sentence scoring was –2.87 dB, with a standard deviation of 0.76 dB. Within-subject variability was illustrated by the standard deviations for each subject's scores, which ranged between 0.96 and 1.79 dB. The mean score and standard deviation for each of the lists across subjects (n=10) are indicated in Figure 1. These means lay between the best performance at –4.13 dB (list 8) and the poorest at –1.40 dB (list 20), a total range of 2.73 dB. Standard deviations for lists ranged between 0.69 and 1.50 dB.

The scores of each subject on each list were also compared to the overall mean of all subjects for all lists. The mean deviations were all within +/- 1.50 dB from the overall mean. The standard deviations from these means varied between 0.69 and 1.50 dB. The variability between lists was further investigated using the Friedman test. Owing to the small sample size (n = 10), this non-parametric test was chosen in favour of a parametric repeated measures analysis of variance (ANOVA). The analysis showed that there were significant differences within the collection of lists ($p < 0.0001$). Additional paired comparisons were performed between lists. Calculated z-values compared to the critical z-value (3.70, $\alpha = 0.05$) showed that average scores obtained for lists 8, 11 and 21 were significantly different from those obtained with a number of other lists. The critical z-value was corrected for the number of comparisons.

Phase IV: Evaluating list equivalence in a reduced bandwidth condition

Method

The method used to attain inter-list equivalence in the current research was the commonly reported method of arranging sentences into phonemically matched lists. Because the phonemic content of the different lists are controlled, it could be inferred that the spectral content of the different lists should be equivalent. If this assumption is true, the different lists should still be of equivalent difficulty even if the bandwidth is reduced by filtering the sentences to eliminate some frequency components. To test this hypothesis, the sentence material was filtered using a low-pass filter with a cut-off frequency of 2000 Hz, and a roll-off slope of 48 dB per octave. This particular filter was selected to resemble to some extent the low-pass filter effect of a high-frequency hearing loss (Stuart et al, 1995; Scott et al, 2001). The use of such a filter does not fully simulate all the effects of a cochlear hearing loss, but was used in this study to test the effect of reduced bandwidth on inter-list equivalence, based on the hypothesis that the difficulty of all the phonemically balanced lists should be equally affected by such a filter.

The filtered sentence materials were presented in the same order to the same normal-hearing subjects that had previously listened to the unfiltered version (Phase III). The same listeners were used in order to enable the researchers to compare exactly the same listener's performance on the unfiltered list with his/her performance on the filtered list. This is because, even with normal-hearing listeners, there are inter-subject differences that could affect this comparison. Also, a possible learning effect was not considered important, as the idea was not to determine whether listeners' performance changed or remained the same with each filtered list, but whether the inter-list differences increased with a reduction in bandwidth. List

equivalence was explored during this phase using the Friedman ANOVA. In addition, the results were compared to the findings of Phase III using the Wilcoxon-rank sum test, which indicates the magnitude and direction of differences for pairs of scores, and tests the significance of the difference between dependent samples (Maxwell & Satake, 2006).

Results

The results obtained during Phase IV are compared to those of Phase III in Table 1. All the values in the table (with the exception of the within-subject standard deviations) are for all lists across all subjects. Within-subject standard deviations were calculated by determining the standard deviation for each subject across all lists, and then calculating the mean of this standard deviation across subjects. Each individual list's scores for Phase III and IV were also compared using the Wilcoxon-rank sum test. It was found that each of the lists showed significant differences ($p<0.05$) between unfiltered (Phase III) and filtered (Phase IV) conditions, indicating that the reduction in bandwidth had a significant effect on the difficulty of the sentence lists.

The inter-list variability as observed in the reduced bandwidth condition (Phase IV) was also investigated. The ANOVA indicated that there were significant differences within the collection of filtered lists ($p<0.0001$), and paired comparisons (corrected for the number of comparisons) revealed that two lists (15 and 21) differed significantly from a number of other lists. These two lists yielded the worst performance (highest SNR required for 50% intelligibility) with an SNR of 2.73 dB and best performance (SNR of -1.20 dB) respectively, and also showed the largest deviations from the overall mean in this experiment (1.93 and –2.01 dB respectively).

After completion of Phase IV, the results from Phases III and IV were used to improve uniformity within the sentence collection. Lists that yielded results significantly different from the others were excluded from the collection. Some lists yielded similar scores in Phase III, but differed significantly in the filtered condition (Phase IV), and the results of both phases were therefore considered in order to improve equivalence. Based on these results, four lists were excluded (lists 8, 11, 15, and 21), as these lists differed significantly from the other lists in terms of performance and/or showed a large deviation (≥ 1 dB) from the overall mean in both the full and reduced bandwidth conditions. Excluding these lists reduced the standard deviation from 0.76 to 0.64 dB, and average within-subject standard deviation from 1.26 to 1.22 dB, leaving a total collection of 18 sentence lists. The average SNR-50 for this final collection of lists (n = 18) was -2.73 dB, with a standard deviation of 0.64 dB.

Phase V: Evaluating test-retest reliability

Method

The final phase of the project was aimed at evaluating the reliability of the test on repeated measures. Twenty young adults (ten female, ten male) aged between 19 and 23 years (mean age = 21.4 years) participated in this project phase. These participants all had hearing thresholds ≤ 15 dB HL at 250, 500, 1 000, 2 000, 4 000 and 8000 Hz, normal otoscopic results, Type A tympanograms indicating normal middle-ear functioning, and no history of otologic disease. The same equipment and test environment described for Phase II were used in this project phase, and the test material used was the phonetically balanced sentence lists as developed and refined during the previous project phases. The adaptive up-down test procedure as described under Phase III was used for this phase as well. Each participant was tested using all 18 lists selected during Phase IV. The order of list presentation was counterbalanced between subjects by starting 18 of the subjects each with a different

list, and picking two randomly selected lists as the first list for the remaining two subjects. All subjects were retested approximately two weeks after their initial test, following the same procedure and list order.

The difference between the mean SNR-50 value for the complete collection of lists (presented to all participants) on the first test and the retest was determined, providing an indication of the test-retest reliability. These differences were analysed using a paired-sample *t*-test, to determine the significance of the difference between test and retest scores across the collection of lists. The lists were compared as a group and not individually, since the equivalence within the collection has already been demonstrated in previous phases, and the clinical application of the test would typically involve testing a listener with different lists to those used during the first evaluation.

Results

The results attained with each list across subjects in the first test were compared to the results attained during the retest. A summary of the results are provided in Table 2. For individual subjects, average performance (across all lists) improved with between 0.37 and 1.41 dB, with an average within-subject improvement of 0.83 dB. The average improvement in performance between test and retest was 0.82 dB (with subjects attaining 50% correct recognition at an SNR that was 0.82 dB worse, i.e. more difficult, during the retest). The standard deviation from this mean was 0.46 dB. Individual lists showed an improvement between the first and second tests of between 0.10 and 1.91 dB across subjects. Note that this value was calculated as the mean of the 18 difference values, that is, the average of each of the 18 lists' difference between test and retest scores. A paired sample *t*-test showed that the difference between test and retest results for all subjects across all lists (20 subjects x 18 lists = 360) was statistically significant (with $t(359) = 10.102$, $p < 0.001$).

DISCUSSION

The combination of translated and original material yielded a sentence collection of equivalent grammatical complexity and naturalness. The characteristics of the sentence content (complete sentences, representative of everyday speech, and free from proverbs, questions, exclamations and proper nouns) ensured that the material received a high rating of naturalness and very few changes or exclusions (13 in total) were necessary to improve this aspect. The grammar rating awarded to each sentence during Phase I was compared to its intelligibility during Phase II, and findings indicated that there was no significant correlation between grammatical complexity and intelligibility in noise. This may be because the sentence collection was relatively uniform in terms of its grammatical complexity (all sentences were at between three- and five-year-old level).

The application of these sentences to normal-hearing listeners during the second project phase provided the means to determine their psychometric slope. Previous studies have used these slopes as an indication of the SNR adjustments needed to equate the intelligibility of the sentences (Nilsson et al, 1994; Vaillancourt et al, 2005; Wong & Soli, 2005; Wong & Huang, 2008), and have found slopes in the range of 9 to 17.9% / dB. In the present study, however, the slopes were used to identify sentences with similar performance under different conditions, which enabled the researchers to exclude sentences that differed significantly from the majority, much like Versfeld et al (2000). Excluding sentences instead of re-scaling intensities reduced the number of subjects and hours of data collection, as previous researchers have sometimes had to conduct up to seven rounds of testing in order to verify the effect of the re-scaling procedure (Nilsson et al, 1994). The average sentence slope of the current study (determined using syllable scoring) was 16.0% / dB, which compares well to the average of other studies. Note that the slope determined in this

way does not apply to the final test format, since the test procedure and scoring method differs. Slopes were used during test development to guide the selection of sentences with equivalent intelligibility in noise. In the final test format, i.e. an adaptive procedure with sentence scoring to obtain the SNR at 50% correct recognition, only the equivalence at the 50% point on the slope is important. In the current project, the final collection of sentences showed a similar degree of equivalence at the 50% point to that reported by previous researchers who used the method of re-scaling intensities (standard deviation in threshold of 0.64 dB across lists, as compared to 0.78 dB reported by Nilsson et al, 1994, and 0.75 dB reported by Wong & Huang, 2008, for example). Should these lists be used in a different test format to that validated in this study, for example to determine the percentage of whole sentences recognised correctly at a specific SNR, the slopes of the lists become an important issue, as different lists may differ in difficulty at points other than the 50% point on the psychometric slope. The only recommended use of the developed material is therefore the adaptive procedure described in Phases III and IV, and other applications of the material would have to be validated before applying these in clinical practice.

The accuracy with which the selected sentences were arranged into phonemically matched lists in the current research compared favourably with the phonetic balance reported in previous studies, as shown in Table 3. In addition, the results attained with these lists in normal-hearing listeners also corresponded with previously reported findings in other languages. These values are compared in Table 4.

The average SNR-50 across subjects (represented in the first column) indicates the average threshold obtained with the final set of 18 lists across the entire group of subjects. The current study's results (average of -2.73 dB) showed greatest correspondence with the findings of Nilsson et al (1994). Perhaps more important

than the absolute values of the thresholds, however, are the standard deviations of these averages, which give an indication of the variability within the collection of lists. Other studies have reported standard deviations ranging from 0.27 to 1.2 dB (with an average of 0.77 dB). The current study found a standard deviation of 0.64 dB, indicating a comparable degree of equivalence across lists.

The within-subject standard deviations for the current lists were found to be slightly larger (1.22 dB) than previously reported deviations, indicating a greater degree of variability between lists for specific subjects. The deviation from the overall mean was also calculated for each list. In the current study, the average deviations of all the lists were found to be within +/- 1.5 dB (ranging between –0.73 and 1.47) from the overall mean. This was slightly larger than deviations reported in a number of other studies (Nilsson et al, 1994; Vaillancourt et al, 2005; Wong & Soli, 2005; Hällgren et al, 2006; Van Wieringen & Wouters, 2008). The larger degree of variability despite the high degree of phonetic balance indicates that list equivalence is affected by factors other than phonetic content.

Nilsson et al (1994) examined the effect of audible bandwidth on the reliability of sentence recognition measurements. Different reductions in bandwidth were tested and it was found that thresholds increased significantly in the reduced bandwidth conditions (Nilsson et al, 1994). Standard deviations did not increase significantly when the 4000 Hz octave band was eliminated, but only when the 2000 Hz octave band was also eliminated (Nilsson et al, 1994). With this bandwidth, standard deviations for five lists in noise increased to about 3 dB. In the current research, a low-pass filter (from 2000 Hz, with a roll-off slope of 48 dB / octave) was used to reduce the bandwidth of the sentence material, and the same groups of subjects used for the initial experimental application of the lists were retested under this condition. The results of this experiment showed an increase in standard deviation

for 22 lists (from 0.76 dB to 0.92 dB), although the deviations were still well below the deviation of 3 dB reported by Nilsson et al (1994). Standard deviations were therefore still close to 1 dB, despite the additional variability between lists caused by the filter. Within-subject standard deviations also increased only slightly (from 1.26 to 1.69 dB).

The fact that the lists were not completely equivalent in either full or reduced bandwidth conditions demonstrates that equivalence in terms of phonetic content does not guarantee list equivalence. Reduced bandwidth could affect test reliability by increasing guessing and response biases (Nilsson et al, 1994). The results of the current research demonstrate that a reduction in audible bandwidth, which is one of many effects that a hearing loss would have on the material, could affect list equivalence despite the fact that the lists are phonemically balanced. This finding corroborates the report by McArdle and Wilson (2006), which demonstrated that not all the sentence lists of the Quick Speech in Noise test (QuickSIN) that are equivalent in listeners with normal hearing are equivalent in listeners with sensorineural hearing loss. It is therefore essential to evaluate list equivalence in hearing impaired populations before assuming that phonetically balanced lists can be used to evaluate individuals with a hearing impairment reliably. In the current study, inter-list equivalence was improved and significant inter-list differences eliminated by excluding lists that differed significantly from other lists in the collection based on results found in listening experiments with full and reduced bandwidth. The question remains, however, what effect a different filter (affecting different frequencies) would have on these lists, and further experimentation to validate inter-list equivalence on listeners with sensorineural hearing impairments could increase the clinical usefulness of the developed test.

The evaluation of the test-retest reliability of the 18 lists selected during Phase IV revealed a significant overall improvement between the first test and a retest of 20 normal-hearing young adults. The difference between the average test and retest score for each list across subjects was always positive, i.e. an improved score during the retest. This improvement between test and retest results could therefore indicate a learning effect over the two test conditions that may be attributed to factors such as memory and familiarisation with the testing procedures (Cameron & Dillon, 2007). Although these differences were found to be statistically significant, they were all below 2 dB. This variability is primarily influenced by the step size used during adaptive testing (Wong et al, 2007). The step size in the current study was 2 dB, and the average difference between test and retest for all lists was therefore smaller than the adaptive step size, indicating an acceptable variability. Furthermore, the improvement noted during this study compares well to the results reported by Hällgren et al (2006) for their test-retest research conducted on the Swedish Hearing in Noise Test. They noted an improvement on the retest condition of 0.77 dB. The results are also comparable to those reported by Cameron and Dillon (2007) who found statistically significant improvements on retest conditions of the LISN-S test (excluding spatial advantage measure) varying from -1.1 dB to 0.1 dB.

CONCLUSION

In the South African context there is a lack of pre-recorded materials for the evaluation of speech perception, especially in languages other than English. The research process described in this report addressed this need by attaining the main aim set for this project, namely the development of a valid and reliable Afrikaans test of sentence recognition thresholds in noise. The validity of the test lies in the fact that the sentences are natural and representative of everyday speech, as rated by native speakers of Afrikaans. In addition, the presence of background noise ensures that the test is conducted in the type of listening situation that commonly occurs in daily

life and constitutes a major challenge for individuals with a hearing impairment, thus providing a valid measure of everyday auditory functioning. The reliability of the test was evaluated by applying the sentence collection to a group of normal-hearing young adults. The results showed a high degree of inter-list equivalence (reliability), and it should therefore be possible to assess this skill repeatedly using the different lists in order to monitor progress or evaluate adjustments made to amplification devices. In addition, the effect of a reduction in bandwidth on inter-list equivalence was evaluated and these results, in conjunction with the results from the first experiment, were used to improve the uniformity of the list collection by eliminating lists with significant differences from other lists in the collection. The test-retest reliability in normal-hearing listeners was evaluated and confirmed during the final project phase. The test developed during this project therefore constitutes a valuable resource to audiologists in South Africa, as it can be used to quantify a common problem in patients with hearing impairment. The developed test in its final format can be obtained from the researchers.

ACKNOWLEDGMENTS

This project was financially supported by the National Research Foundation of South Africa. The authors would like to express gratitude to Celeste Botha and Chantelle Cater for their assistance with the data collection process.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES

Bellis, T.J. 2003. Auditory processing disorders: It's not just kids who have them. *The Hearing Journal,* 56, 10-18.

Bench, J. & Bamford, J. 1979. *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children.* London: Academic Press.

Bevilacqua, M.C., Banhara, M.R., Da Costa, E.A., Vignoly, A.B. & Alvarenga, K.F. 2008. The Brazilian Portuguese Hearing in Noise Test. *Int J Audiol,* 47, 364-365.

Boersma, P. & Weenink, D. 2006. *Praat: doing phonetics by computer* (Version 4.3.14) [Computer program]. Retrieved June 20, 2006, from http://www.praat.org/

Buitendag, M.M. 1994. Die opstel en standaardisering van 'n Afrikaanse Reseptiewe Woordeskattoets. Unpublished M Communication Pathology research report. Department of Communication Pathology, University of Pretoria.

Cameron, S. & Dillon, H. 2007. Listening in spatialized noise-sentences test (LISN-S): test-retest reliability study. *Int J Audiol,* 46, 145-153.

Carstens, W.A.M. 2003. *Norme vir Afrikaans: Enkele riglyne by die gebruik van Afrikaans* (4th ed.). Pretoria: Van Schaik Uitgewers.

Cekic, S. & Sennaroglu, G. 2008. The Turkish Hearing in Noise Test. *Int J Audiol,* 47, 366-368.

Central Statistical Service. 1995. *Provincial Statistics 1995* (report number 00-90-07). Pretoria: Central Statistics.

Hällgren, M., Larsby, B. & Arlinger, S. 2006. A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *Int J Audiol,* 45, 227-237.

Huarte, A. 2008. The Castilian Spanish Hearing in Noise Test. *Int J Audiol,* 47, 369-370.

Killion, M.C. & Niquette, P.A. 2000. What can the pure-tone audiogram tell us about a patient's SNR loss? *The Hearing Journal, 53*(3), 46-53.

Killion, M.C., Niquette, P.A. & Gudmundsen, G.I. 2004. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am,* 116, 2395-2405.

Kollmeier, B. & Wesselkamp, M. 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J Acoust Soc Am,* 102, 2412-2421.

Le Roux, T.H. & Pienaar, P. de V. 1976. *Uitspraakwoordeboek van Afrikaans.* Pretoria: J.L. van Schaik.

Lolov, S.R., Raynov, A.M., Boteva, I.B. & Edrev, G.E. 2008. The Bulgarian Hearing in Noise Test. *Int J Audiol,* 47, 371-372.

Lutman, M.E. 1997. Speech tests in quiet and noise as a measure of auditory processing. In M. Martin (ed.), *Speech Audiometry* (2nd ed.) (pp. 63-73). London: Whurr Publishers Ltd.

Luts, H., Boon, E., Wable, J. & Wouters, J. 2008. FIST: A French sentence test for speech intelligibility in noise. *Int J Audiol,* 47, 373-374.

Maxwell, D.L. & Satake, E. 2006. *Research and Statistical Methods in Communication Sciences and Disorders.* Canada: Thomson Delmar Learning.

McArdle, R.A. & Wilson, R.H. 2006. Homogeneity of the 18 QuickSIN™ Lists. *J Am Acad Audiol,* 17, 157-167.

Moon, S.K., Kim, S.H., Mun, H.A., Jung, H.K., Lee, J., Choung, Y. & Park, K. 2008. The Korean Hearing in Noise Test. *Int J Audiol,* 47, 375-376.

Myhrum, M. & Moen, I. 2008. The Norwegian Hearing in Noise Test. *Int J Audiol,* 47, 377-378.

Nilsson, M.J., 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, 1085-1099.

Plomp, R. & Mimpen, A.M. 1979. Improving the Reliability of Testing the Speech Reception Threshold for Sentences. *Audiology,* 18, 43-52.

Quar, T.K., Mukari, S.Z.M.S., Wahab, N.A.A., Razak, R.A., Omar, M. & Maamor, N. 2008. The Malay Hearing in Noise Test. *Int J Audiol,* 47, 379-380.

Scott, T., Green, W.B., & Stuart, A. 2001. Interactive Effects of Low-Pass Filtering and Masking Noise on Word Recognition. *J Am Acad Audiol*, 12, 437-444.

Shiroma, M., Iwaki, T., Kubo, T. & Soli, S. 2008. The Japanese Hearing in Noise Test. *Int J Audiol,* 47, 381-382.

Smits, C., Kramer, S.E., & Houtgast, T. 2006. Speech Reception Thresholds in Noise and Self-Reported Hearing Disability in a General Adult Population. *Ear Hear,* 27, 538-549.

Soli, S.D. & Wong, L.L.N. 2008. Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int J Audiol,* 47, 356-361.

Statistics South Africa. 2001. *Census 2001 Key Results.* Retrieved September 8, 2006 from http://www.statssa.gov.za/

Stuart, A., Phillips, D.P. & Green, W.B. 1995. Word recognition performance in continuous and interrupted broad-band noise by normal-hearing and simulated hearing-impaired listeners. *The American Journal of Otology*, 16, 658-663.

Theunissen, M., Swanepoel, D. & Hanekom, J. 2009. Sentence recognition in noise: variables in compilation and interpretation of tests. *Int J Audiol,* 48, 743-757.

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A. Soli, S.D. & Giguère, C. 2005. Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *Int J Audiol*, 44, 358-369.

Van Wieringen, A. & Wouters, J. 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and The Netherlands. *Int J Audiol,* 47, 348-355.

Versfeld, N.J., Daalder, L., Festen, J.M. & Houtgast, T. 2000. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am*, 107, 1671-1684.

Wong, L.L.N & Huang, V. 2008. The Taiwanese Mandarin Hearing in Noise Test. *Int J Audiol,* 47, 391-392.

Wong, L.L.N & Soli, S.D. 2005. Development of the Cantonese Hearing In Noise Test. *Ear Hear,* 26, 276-289.

Wong, L.L.N., Soli, S.D., Liu, S., Han, N. and Huang, M. 2007. Development of the Mandarin Hearing in Noise Test (MHINT). *Ear Hear,* 28, Supplement, 70S-74S.

TABLES

Table 1: Comparison of results from Phase III and IV, for all 22 lists

Variables	*Phase III Full bandwidth (dB SNR)*	*Phase IV Reduced bandwidth (dB SNR)*	*Difference (dB SNR)*
Average	-2.87	0.81	3.68
Standard deviation	0.76	0.92	0.16
Minimum	-4.13	-1.2	2.93
Maximum	-1.4	2.73	4.13
Range (maximum - minimum)	2.73	3.93	1.2
Within-subject standard deviation	1.26	1.69	0.43

Table 2: Comparison of results from first test and retest during Phase V. Values shown are for all 18 lists across all 20 participants of this phase (Std dev = standard deviation)

	First test (dB SNR)	*Retest (dB SNR)*	*Mean difference across lists (dB SNR)*
Average	-2.6	-3.35	0.82
Std dev	0.71	0.78	0.46
Minimum	-4.13	-4.5	0.1
Maximum	-1.2	-2.03	1.91

Table 3: Accuracy of phonetic balancing of developed lists (Afrikaans lists) with previous reports

Variables	*Afrikaans lists*	*Nilsson et al (1994)*	*Vaillancourt et al (2005)*	*Wong & Soli (2005)*	*Hällgren et al (2006)*
% phoneme counts where error = 0	35%	30-35%	35%	<15%	
% of counts within +/- 1 phoneme	83%	68%	75%	+/- 61%	70%

Table 4: Comparison of Phase III findings to previous reports (std dev = standard deviation)

Authors	*Average SNR-50 across subjects (dB)*	*Std dev from average (dB)*	*Within-subject std dev (dB)*
Plomp & Mimpen (1979)	x	x	0.9
Nilsson et al (1994)	-2.9	0.78	1.13
Kollmeier & Wesselkamp (1997)	-6.2	0.27	x
Versfeld et al (2000)	-4.1	0.56	1.07
Vaillancourt et al (2005)	-3.3	0.5	1.1
Wong & Soli (2005)	-3.9	1.0	1.8
Hallgren et al (2006)	-3	1.1	x
Wong et al (2007) MHINT-M	-4.3	0.62	0.89
Wong et al (2007) MHINT-T	-4.0	0.94	0.75
Van Wieringen & Wouters (2008)	-7.8	1.2	1.17
Bevilacqua et al (2008)	-4.6	0.8	1.2
Cekic & Sennaroglu (2008)	-3.9	0.9	1.0
Huarte (2008)	-3.6	1.2	x
Lolov et al (2008)	-4.0	1.5	0.6
Moon et al (2008)	-3.3	1.0	1.9
Myhrum & Moen (2008)	-3.2	1.0	0.9
Quar et al (2008)	-4.7	0.8	1.3
Shiroma et al (2008)	-5.3	1.4	x
Average	*-4.24*	*0.92*	*1.12*
Minimum	*-7.8*	*0.27*	*0.6*
Maximum	*-2.9*	*1.5*	*1.9*
Current findings	*-2.73*	*0.64*	*1.22*

FIGURES

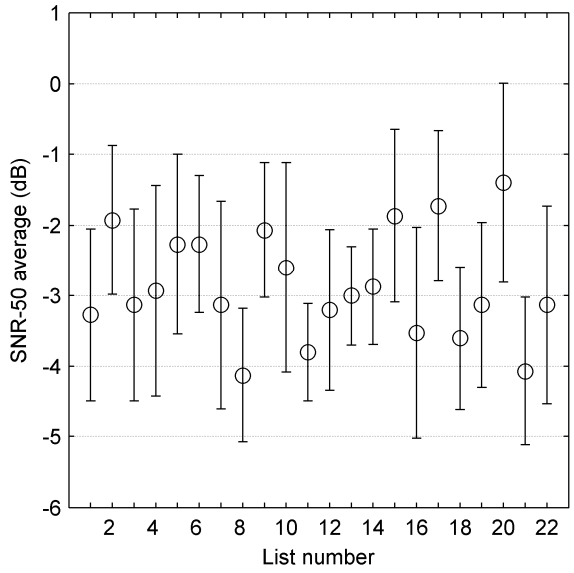

Figure 1: Mean SNR-50 across subjects (n=10) for each of the 22 lists. Error bars indicate +/- one standard deviation for each list.