A Clinical Evaluation of an Automated Technique for Estimating Speech Reception Thresholds[1]

Évaluation clinique d'une méthode de mesure automatique du seuil d'intelligibilité de la parole

Donald G. Jamieson and *Margaret F. Cheesman*
Hearing Health Care Research Unit
Department of Communicative Disorders
The University of Western Ontario, London, Ontario

Hollis Corbin
Otologic Function Unit
Mt. Sinai Hospital, Toronto, Ontario

Key words: speech audiometry, speech reception threshold

Abstract

This article describes an implementation of Cheesman's (1992) procedure for estimating speech reception thresholds (SRTs) of individual hearing-impaired listeners, and evaluates the procedure in a clinical setting. The procedure uses a closed set with six spondaic words as targets; the level of these speech signals is adjusted from trial to trial, using an adaptive tracking procedure, so that the SRT corresponding to the desired identification probability level can be determined rapidly. Response alternatives are presented on a computer monitor, and the listener responds using a computer mouse. Responses are recorded automatically by the computer, and the signal level for the next trial is adjusted by the computer, according to the adaptive algorithm. Test-retest reliability is extremely high, and the test results are well correlated with pure-tone, air-conduction thresholds (r=.86 with PTA and r=.82 with the best threshold at any audiometric frequency between 500 and 4000 Hz). Adaptive SRT scores also correlate well with SRT scores measured by conventional clinical procedures (r=.82). Finally, the test can be made still more efficient by modifying the stopping rule for the adaptive algorithm; thresholds calculated using fewer trials were strongly related to those of the full test.

Résumé

Suit une description de l'application de la méthode de Cheesman (1992) permettant de mesurer le seuil de réception de la parole (SRT) des auditeurs malentendants, et de l'évaluation de cette méthode en milieu clinique. Cette méthode utilise un jeu de six mots spondaïques comme cible. On corrige le niveau des signaux vocaux d'un essai à l'autre grâce à une technique de contrôle adaptive en vertu de laquelle on parvient à déterminer rapidement le SRT en fonction du degré d'identification probable souhaité. Les diverses réponses possibles apparaissent sur un moniteur et l'auditeur indique la bonne au moyen d'une souris. Les réponses sont enregistrées automatiquement par l'ordinateur qui ajuste le niveau du signal à l'essai suivant, conformément à l'algorithme d'adaptation. On note un très haut degré de fiabilité entre l'épreuve initiale et l'épreuve de contrôle. Les résultats des essais présentent une bonne corrélation avec le seuil de conduction aérienne des sons purs (r=0,86 avec conduction aérienne des sons purs et r=0,82 avec le meilleur seuil à une fréquence audiométrique de 500 à 4000 Hz). Le SRT corrigé présente également une bonne corrélation avec le SRT mesuré par les méthodes cliniques habituelles (r=0,82). Enfin, on peut accroître encore plus l'efficacité de l'essai en modifiant la règle d'arrêt de l'algorithme d'adaptation; les seuils calculés avec un plus petit nombre d'essais étaient fortement liés à ceux obtenus avec le test intégral.

While speech intelligibility testing remains a part of the routine battery of clinical audiometric procedures, the conventional tests that can be administered relatively quickly — for example, those using 25 monosyllabic words or live-voice speech reception threshold tests — are so insensitive that they can differentiate only among the most extreme differences in performance (Thornton & Raffin, 1978). Unfortunately, the tests that are more reliable, in particular those using pre-recorded, calibrated stimuli and those having more test items, such as multiple-list phonetically-balanced tests, require considerable time to administer and are unsuited to automated stimulus delivery and scoring procedures. As a consequence, they are not particularly well-suited to routine clinical use.

For some time, our group has been involved in developing advanced measures of speech intelligibility which are time-efficient, reliable, valid and, wherever possible, automated, to minimize the possibilities for human error and make the best possible use of the clinic or laboratory time (e.g., Cheesman, 1992; Cheesman, Lawrence, & Appleyard, 1992; Jamieson, Dell'Orletta, & Ramji, 1988). Cheesman (1992) described an adaptive speech reception threshold (ASRT) testing procedure that was efficient (i.e., relatively fast) and highly reliable (Cheesman, Jamieson, Seewald, & Gagné, 1990). Levitt's (1971) adaptive tracking paradigm was applied to obtain an efficient estimate of a listener's

threshold. Stimulus presentation, level adjustment and scoring was controlled by computer. The procedure is adaptive in the sense that the level of the speech is determined by the correctness of the listener's previous response(s). Data are collected and scored automatically, and without prejudice or error, by computer. This approach to testing is consistent with that of several previous investigations that have assessed the effects of noise on speech perception in normal hearing and hearing-impaired listeners and that have compared the perception of speech processed by different hearing aid circuitry (e.g., Dirks, Morgan, & Dubno, 1982; Plomp & Mimpen, 1979; Van Tasell, Larsen, & Fabry, 1988; Van Tasell & Yanz, 1987).

Recent reports have demonstrated the utility of the SRT both for assessment and as a measure of the outcomes of aural rehabilitation interventions (e.g., Jamieson & Brennan, 1992; Lee & Humes, 1993; Moore, Lynch, & Stone, 1993; Nilsson, Soli, & Sullivan, 1993). The advantages of the adaptive SRT approach as determined in a laboratory situation can be expected to extend to the clinical setting (cf., Lutman, 1987). For these reasons, we were interested in evaluating the potential clinical utility of Cheesman's adaptive speech reception threshold estimation procedure. (The reader is referred to Cheesman (1992) for a description of the development and initial laboratory evaluation of the adaptive testing procedure.) The present article describes an adaptation of the testing procedure for clinical purposes, and an evaluation of that procedure in an audiology clinic.

Method

Subjects

Subjects were 50 adults aged 23 to 71; all were regular clients of the Mount Sinai Hospital Otologic Function Unit (OFU) in Toronto who agreed to participate in this additional test during their audiological examination. No attempt was made to prescreen clients on the basis of hearing loss or other factors. Analyses are based on the 40 clients for whom clinical audiometry data were complete and for whom two ASRT measures were obtained on at least one ear. However, clinical impressions are provided on the complete experience with all 50 clients. The data include 27 subjects for whom both ears were tested, and 13 subjects for whom only one ear was tested.

Equipment and Stimuli

Pure-tone testing was assessed with a clinical audiometer (Lucas/Grason-Stadler GSI-16). Signals for clinical speech reception threshold measurement were replayed from a compact disk player (Sony CDP-690), and into the audiometer through the tape input channel, so that the audiometer could be used to adjust the signal level manually. Stimuli were presented over supra-aural earphones (Telephonics TDH 50P).

The ASRT procedure used digital recordings of the six spondees (spoken by a male voice) described by Cheesman (1992). These were converted to 12-bit samples at a 20 kHz sample rate using the CSRE 4.0 software (Jamieson, Ramji, Kheirallah, & Nearey, 1992). Spondees were reproduced via a digital-to-analog converter (DataTranslation 2801A), low-pass filtered (Krohn Hite 3700) at 8 kHz, attenuated (TDT PA3 attenuator), amplified (Realistic SA-150), and presented to listeners over supra-aural earphones (Telephonics TDH 39P). All stimuli were presented in quiet.

Clinical Tests

All subjects received routine clinical audiometry in the OFU, including air and bone conduction thresholds and speech reception thresholds. All testing was done by OFU audiologists. Air conduction thresholds were estimated using a modified ascending-descending procedure with a 5-dB step size, to seek the point at which listeners were correct on 3 of 5 trials. A similar procedure was used to estimate clinical SRT (CSRT). The level required for listeners to be correct on 3 of 5 trials was estimated for spondees spoken by a female talker with a 4 second interstimulus interval (replayed in sequence from Track 2 of the VA's Compact Disk 1; Veterans Administration, 1989). Spondee level was adjusted in 5-dB steps.

ASRT Procedure

Listeners were tested while seated in a double-walled sound-attenuating booth (Industrial Acoustics Corporation), in a different room from that in which clinical testing occurred. On each trial, 1 of the 6 spondees was selected at random and presented at the level specified for that trial. Following each stimulus presentation, the response alternatives were displayed on the screen of a computer monitor placed immediately in front of the listener. Listeners used a computer mouse to move a pointer to indicate which word on the screen they had heard and then pushed a button to select their response. After a response was made, the next word was presented automatically. Each response made by a listener was either correct (i.e., the word presented was selected) or incorrect. Levitt's (1971) adaptive tracking algorithm was used to track the 70.7% correct response level. The signal level was decreased after each two correct responses (to make the task more difficult and move "down the psychometric function") and was increased after each

incorrect response (to make the task somewhat easier, moving "up the psychometric function"). By a series of approximations, this method converges on the 70.7% correct point. Each change in the direction of level adjustment is termed a "reversal"; our procedure continued until a total of 11 reversals had occurred. The first three such reversals were discarded and the mean then taken of the next eight reversals.

Words were presented initially at a level well above threshold. The level decreased in 10-dB steps until an error was made. Subsequent adjustments were made in 1.5-dB steps. Speech stimuli were presented until the speech level changes had reversed direction 10 times. The speech levels of the last eight reversals were averaged to obtain a threshold estimate and a measure of response variability. For each listener, two separate ASRT estimates were obtained for each ear.

Test Application

A typical session began by orienting the listener to the task in a 1 to 3 trial practice session during which each of the six spondees were presented once at a clearly suprathreshold level. This practice session permitted subjects to become familiar with the words and to become comfortable with the mouse and monitor response system.

Results and discussion

Clinical Impressions

Most listeners had little difficulty with any aspect of the task. A few had difficulty seeing the small mouse cursor against the monochrome screen, but all were able to perform the task following practice. To address this issue, we now use a color monitor and have enlarged our cursor display. A few listeners had difficulty controlling the mouse, requiring additional practice; one client was unable to perform the task because of this difficulty.

The ASRT procedure is intended to be highly sensitive so that relatively small differences in performance can be detected. As Figure 1 shows, test-retest reliability was remarkably high (r=.98 for test vs retest ASRT estimates). The mean threshold for the first and second estimates of ASRT were 19.1 and 18.4 dB HTL (standard deviations 13.5 and 13.7 dB) respectively. The mean test-retest difference in threshold was 0.6 dB, with a standard deviation of 2.4 dB. This is consistent with Cheesman's report that the standard deviation of the measurement was routinely $\leq$ 3 dB within a run, and that two successive estimates typically fell within 3 dB of each other.

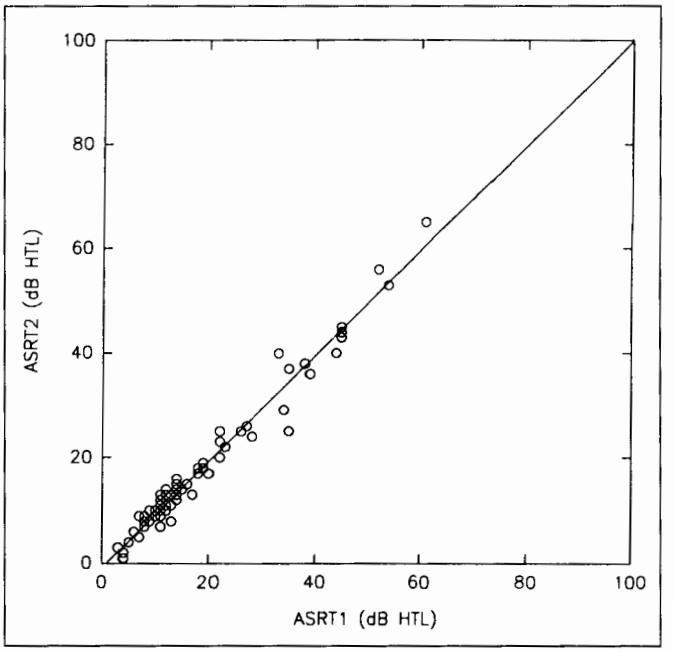

Figure 1. Relation between initial speech reception threshold estimates (ASRT1) and subsequent (retest) estimates (ASRT2), for individual ears.

One typical clinical application of SRT scores is to confirm the results of pure-tone audiometry. As Figure 2 shows, the initial ASRT scores (i.e., ASRT1, the first estimate obtained on a given ear) correlate highly with the average pure-tone hearing threshold at 500, 1000, and 2000 Hz (r=.82) and with the best threshold at any audiometric frequency at octave steps between 500 and 4000 Hz (r=.82). We were also interested in the relation between the ASRT1 scores and the SRT estimates obtained by the clinicians (who had previous knowledge of the pure-tone thresholds). Figure 3 shows that the agreement between these measures was quite good (r=.83 for ASRT1 estimate vs the SRT estimated by conventional clinical procedures).

The correlations with pure-tone thresholds and those with the clinical SRT are particularly impressive in view of the fact that the ASRT values were obtained in a separate room and with different equipment; known sources of variability, including the removal and replacement of supra-aural headphones and the different speech materials (male vs female voices) worked to reduce the strength of the relationship. The ASRT1 vs. CSRT scatterplot shows that the CSRT and ASRT1 differed by 10 dB or more in 11 cases. The ASRT1 exceeded the CSRT in only 3 of these cases. Possible sources for these differences may be the differential sensitivity of some listeners to the speech materials used in the two tests, particularly when listeners had steeply-sloping high-frequency hearing loss, and the clinicians' expectations and knowledge of the pure-tone audiometry results.

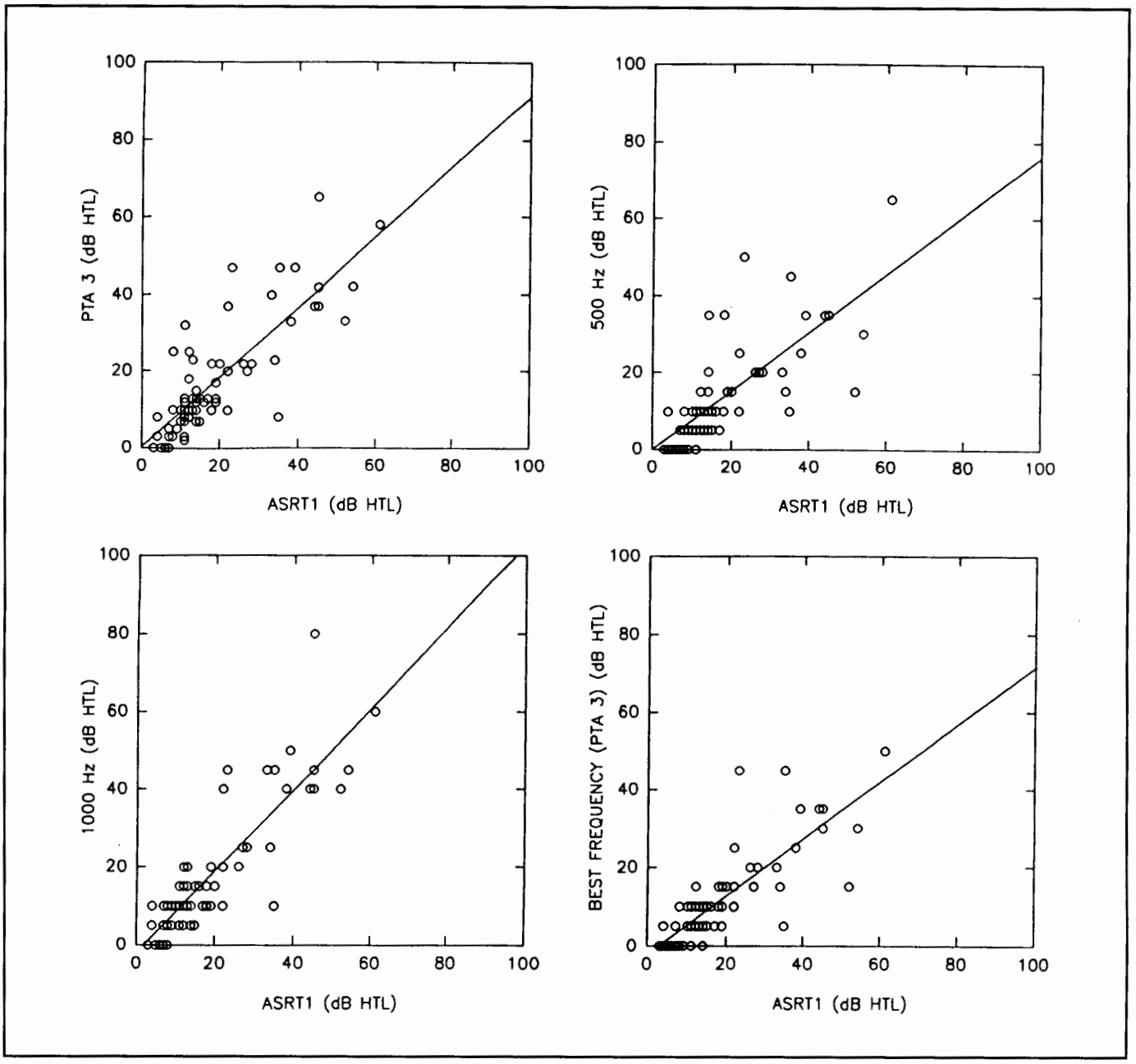

Figure 2. Relation between initial speech reception threshold estimates (ASRT1) and pure-tone, air conduction threshold estimates for individual ears. Results are displayed for PTA (i.e., mean threshold at 500, 1000 and 2000 Hz; upper left panel), for 500 Hz threshold (upper right panel), for 1000 Hz threshold (lower left panel), and for threshold at the frequency with the best residual hearing; lower right panel).

Alternative Adaptive Stopping Rules

In order to increase the efficiency of the adaptive procedure, we examined the effect of altering the stopping rules. We compared four alternative rules by recalculating estimates from the data obtained using the eight-reversal criterion; Figure 4 displays the results for averages based on just two, four, or six (versus eight) reversals. It can be seen that the number of reversals could be reduced from eight to either six or four, without serious estimation error (r=.999 for six vs eight reversals, and r=.998 for four vs eight reversals, respectively). Indeed, in many instances, even two reversals will provide a useful estimate (r=.992 for two vs eight reversals). Such reductions have important consequences for test duration; we found that the mean number of trials to reach the eight-reversal criterion was reduced by 15%, 29% and 46% respectively, for the six-, four-, and two-reversal criteria.

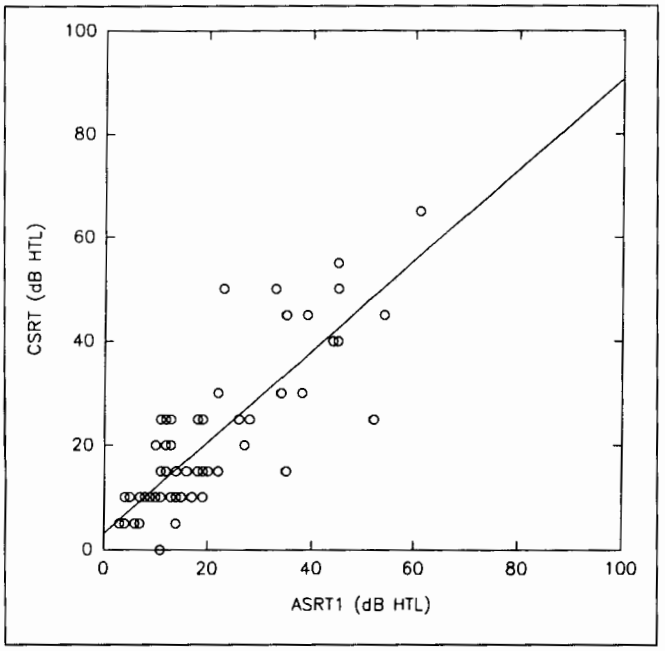

Figure 3. Relation between initial speech reception threshold estimates obtained using the adaptive testing procedure (ASRT1) and the speech reception threshold estimates obtained initially using the conventional OFU clinical procedure (CSRT).

General Discussion

The present modification of Cheesman's (1992) adaptive speech reception threshold procedure appears to offer several advantages for use in an audiology clinic. Stimulus presentation and response collection are automated to ensure that the algorithm is implemented precisely, to minimize the possibilities for error and bias, and to ensure that the audiologist's time is used efficiently. The test is easy for subjects to learn and do, it is highly reliable, and the results are highly correlated to pure-tone thresholds and to SRT scores obtained using more conventional manual testing. Because of the computerized response mode, listeners must be literate

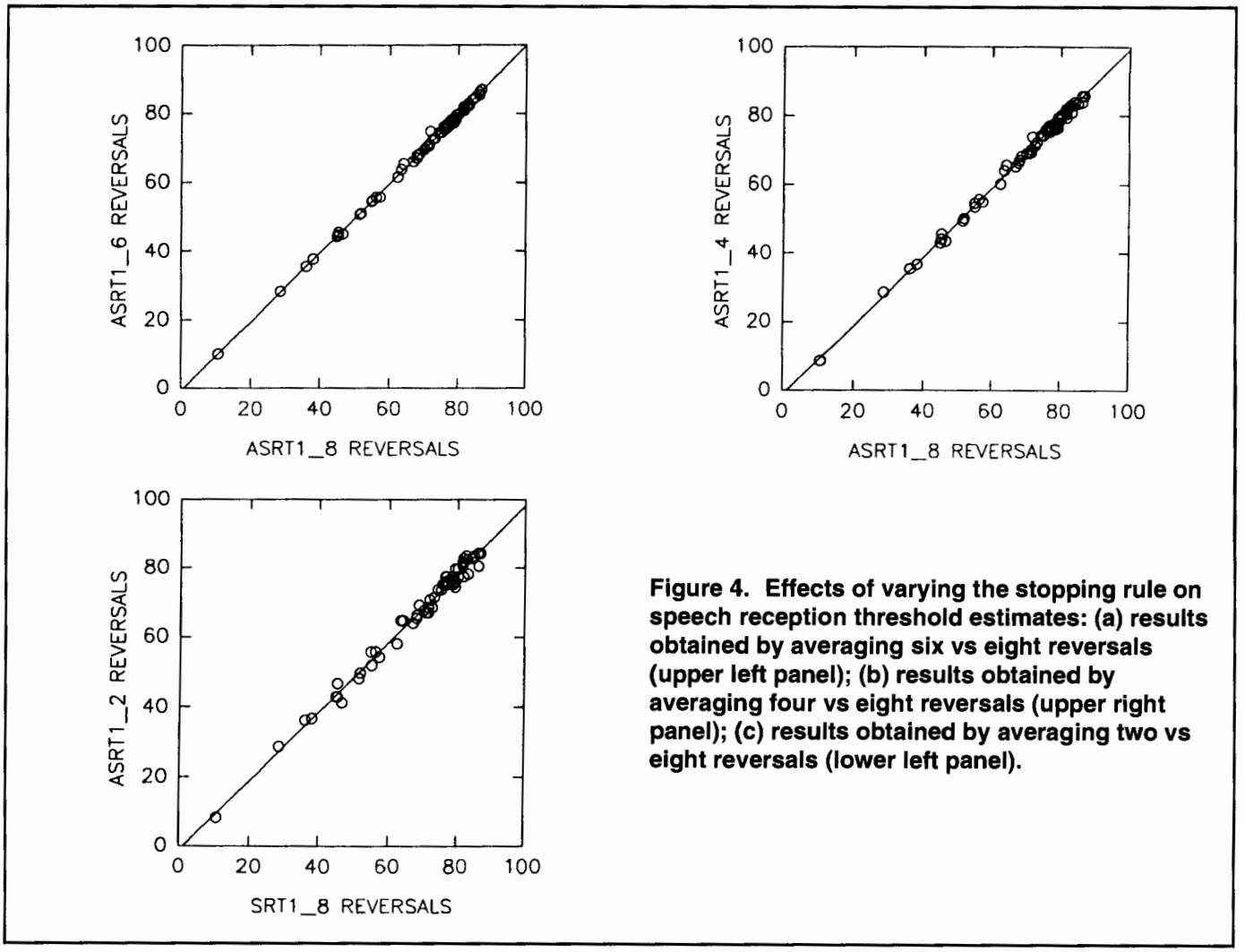

Figure 4. Effects of varying the stopping rule on speech reception threshold estimates: (a) results obtained by averaging six vs eight reversals (upper left panel); (b) results obtained by averaging four vs eight reversals (upper right panel); (c) results obtained by averaging two vs eight reversals (lower left panel).

and have moderately good corrected vision and manual dexterity[2]. Subjects of wide-ranging ages and educational backgrounds typically have little difficulty with the task when they are given a brief opportunity to practice. Consequently, the approach described here should find application in a variety of clinical situations.

References

Cheesman, M.F. (1992). An automated technique for estimating speech reception thresholds in multi-talker babble. *Journal of Speech-Language Pathology and Audiology, 16,* 1-5.

Cheesman, M.F., Jamieson, D.G., Seewald, R. C., & Gagné, J.-P. (1990). Efficient and reliable estimation of speech reception thresholds. *Journal of Speech-Language Pathology and Audiology, 14,* 61.

Cheesman, M.F., Lawrence, S., & Appleyard, A. (1992). Prediction of performance on a nonsense syllable test using the Articulation Index. In Ohala, J.J., Nearey, T.M., Derwing, B.L., Hodge, M.M., and Wiebe, G.E. (Eds.), *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing.* Edmonton: University of Alberta, 1123-1126.

Dirks, D.D., Morgan, D.E., and Dubno, J.R. (1982). A procedure for quantifying the effects of noise on speech recognition. *Journal of Speech and Hearing Disorders, 47,* 114-123.

Jamieson, D.G., and Brennan, R.L. (1992). Evaluation of speech enhancement strategies for normal and hearing impaired listeners. *Proceedings of European Speech Communication Association Conference on Speech Processing Under Adverse Conditions.* 155-163.

Jamieson, D.G., Dell'Orletta, D., & Ramji, K. (1988). Towards the automated assessment of speech intelligibility: Some results from the amplification research group. *Human Communication Canada, 12,* 45-47.

Jamieson, D.G., Ramji, K., Kheirallah, I., & Nearey, T.M. (1992) CSRE: The Canadian speech research environment. In Ohala, J.J., Nearey, T.M., Derwing, B.L., Hodge, M.M., and Wiebe, G.E. (Eds.) *ICSLP 92 Proceedings: 1992 International Conference on Spoken Language Processing,* Edmonton: University of Alberta, 1127-1130.

Lee, L.W., and Humes, L.E. (1993). Evaluating a speech-reception threshold model for hearing impaired listeners. *Journal of the Acoustical Society of America, 93,* 2879-2885.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Lutman, M.E. (1987). Speech tests in quiet and in noise as a measure of auditory processing. In Martin, M. (Ed.). *Speech Audiometry.* Whurr Publishers, London.

Moore, B., Lynch, C. & Stone, M.A. (in press). Effects of the fitting parameters of a two-channel compression system on the intelligibility of speech in quiet and in noise. *British Journal of Audiology.*

Nilsson, M., Soli, S.D. & Sullivan, J. (1993). *Development of the Hearing in Noise Test for measurement of speech reception thresholds in quiet and in noise.* Los Angeles, CA: House Ear Institute.

Plomp, R., and Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18,* 43-52.

Thornton, A., and Raffin, M. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research, 21,* 507-518.

Van Tasell, D.J., Larsen, S.Y., & Fabry, D.A. (1988). Effects of an adaptive filter hearing aid on speech recognition in noise by hearing-impaired subjects. *Ear and Hearing, 9,* 15-21.

Van Tasell, D.J., and Yanz, J.L. (1987). Speech recognition threshold in noise: Effects of hearing loss, frequency response, and speech materials. *Journal of Speech and Hearing Research, 30,* 377-386.

Veterans Administration (1989). *Speech recognition and identification materials disk 1.* (compact disk). Audiology Section, VA Medical Center, Long Beach, California.

Author Notes

[1]This project was supported by grants from the Ontario Ministry of Health, Unitron Industries Limited and URIF.

[2]In another component of the project, we implemented an adaptive SRT procedure for use with children. In this task, the children point to the picture on the screen which corresponds to the object they heard named (e.g., "toothbrush"). This approach has worked well with children who have not yet learned to read.

We are grateful to L. Cornelisse for assistance with calibration and project management, to C. Mandarino and K. Ramji for technical and programming support, to V. Masterson for assistance in data analysis, and to L. Kieffer for assistance in report presentation.

Address all correspondence to: Dr. D.G. Jamieson, Hearing Health Care Research Unit, Elborn College, The University of Western Ontario, London, ON N6G 1H1.