An Automated Technique for Estimating Speech Reception Thresholds In Multi-talker Babble

Technique automatisée d'évaluation du seuil de réception de la parole en présence de bruits de fond

Margaret F. Cheesman
Hearing Health Care Research Unit
University of Western Ontario

Key words: computer, speech reception threshold

Abstract

This paper describes a procedure to estimate the speech-reception thresholds (SRTs) of individual hearing impaired listeners in a rapid, yet highly reliable manner. The test was developed as part of a test battery to evaluate and compare hearing aid performance, but has potential for use in clinical audiometry. The test uses a set of six spondaic words, selected to be homogeneous with respect to the probability of being identified by the listener in a given level of multi-talker babble. The level of the speech signal 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. On each trial, the response alternatives are presented on the screen of a computer monitor, and the listener makes a response using a computer mouse. The listener's response is recorded automatically by the computer, and the signal level for the next trial is then adjusted by the computer, according to rules specified in the adaptive algorithm. The test has been applied successfully to evaluate hearing aid circuitry with young and elderly hearing impaired listeners.

Résumé

Le présent document décrit une méthode d'evaluation rapide et très fiable du seuil du réception de la parole pour les personnes ayant des troubles auditifs. Le test a été conçu pour faire partie d'un ensemble de tests visant à évaluer et à comparer le rendement des prothèses auditives, mais il peut également être utilisé en audiométrie clinique. Le test utilise un ensemble de six spondées, choisies de façon homogène quant à la probabilité d'être identifié par un auditeur en présence de bruits de fond. Le niveau du signal sonore est ajusté d'un essai à l'autre, en se servant d'une méthode de contrôle adaptée pour que le seuil correspondant au niveau probable d'identification désiré puisse être déterminé rapidement. Lors de chaque essai, des choix de réponses apparaissent à un écran d'ordinateur, et la personne qui écoute se sert d'une souris pour donner sa réponse. La réponse est enregistrée automatiquement par l'ordinateur, et le niveau du signal sonore est alors ajusté par l'ordinateur pour l'essai suivant, selon les règlements prescrits par l'algorithme adaptatif. Le test a été utilisé avec succès pour évaluer les circuits des prothèses auditives des handicapés auditifs jeunes et vieux.

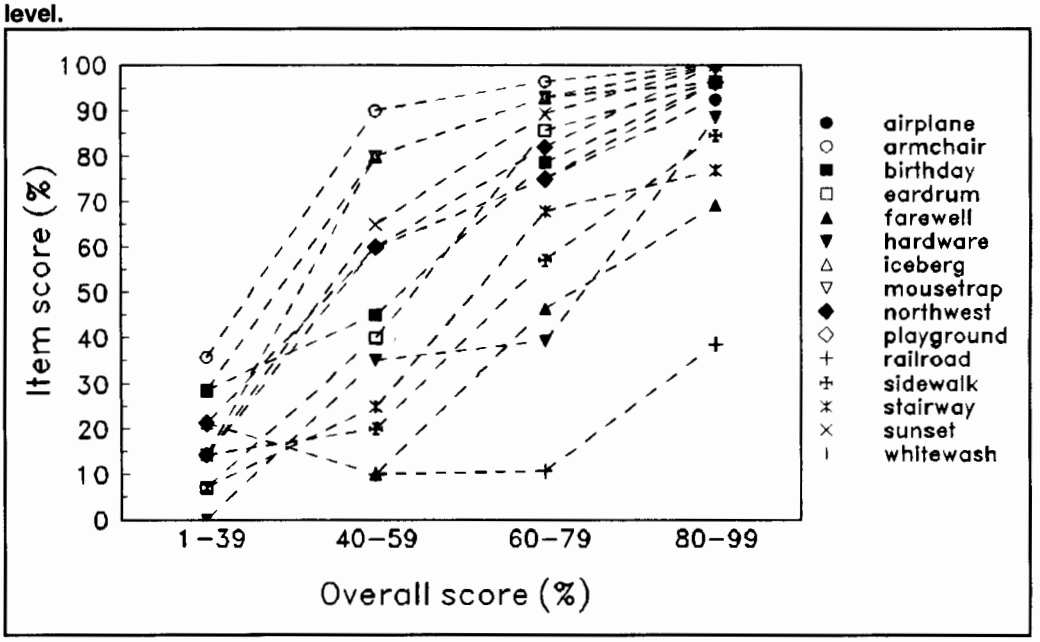
This paper reports one aspect of a project to develop procedures to assess amplification devices along both acoustic and perceptual dimensions. Such assessments require efficient (that is, relatively fast), yet highly reliable speech testing materials to quantify differences among hearing aids. Both threshold and supra-threshold measures of intelligibility are needed.

While many speech audiometry tests are available, most of these were developed for very specific, clinical applications (Olsen & Matkin, 1979, provide an overview of the objectives of speech audiometry; alternative approaches to speech audiometry and contemporary procedures used in various countries are reviewed by Martin, 1987). There are tests that can be administered relatively quickly—for example, those using 25 monosyllabic words or live-voice speech reception threshold tests—but they are too insensitive to differentiate among all but the most extreme differences in performance (Thornton & Raffin, 1978). Unfortunately, tests that are more reliable, in particular those using pre-recorded, calibrated stimuli and those having more test items, such as multi-list PB tests, require considerable time to administer and are unsuited to automated stimulus delivery and scoring procedures. Therefore, they are not particularly well-suited to being included in a battery of tests.

Practice effects are a further concern in evaluation because listeners typically are tested under a variety of amplification conditions; for this reason, test materials must be such that listeners show little or no improvement with practice at the task. Tests must also be easy to explain to listeners and learned readily. Finally, it is desirable to have a testing procedure automated as fully as possible in order to make the test as efficient as possible, to minimize the possibility for human error, and to make the best possible use of the clinical researcher's time.

To meet these various demands, a modified speech reception threshold (SRT) test was developed that uses an

Figure 1. Scores obtained for 15 spondaic words as a function of the overall performance level.



adaptive tracking paradigm and is scored automatically by computer. This test was designed to add a threshold measure of speech reception to a battery of speech intelligibility tests that included suprathreshold measures of word and phoneme discrimination, and listener judgements of the intelligibility and quality of speech, all of which are used for making comparisons among hearing aid circuits.

The adaptive SRT procedure uses the modified adaptive tracking algorithm developed by Levitt (1971). This procedure is adaptive in the sense that the level of the speech is determined by the correctness of the listener's previous response(s). 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).

One prerequisite for optimizing the efficiency of the adaptive tracking algorithm is to ensure that test items are homogeneous with respect to their difficulty at all performance levels that may be covered during an adaptive track. This means not only that the audibility thresholds (usually defined as the level at which 50% of the words would be correctly identified) must be similar for all items, but also that all other points on the performance-intensity function must be the same for all items.

Lists of test items that satisfy the former criterion, when the listening task is conducted in quiet, have been published and used extensively (e.g., ASHA, 1988). However, no list of test items that satisfy the latter criterion exist, nor is there a list of test items that have the same audibility thresholds when listening in babble noise. This is an important concern given the long-term objectives of the present work—the comparative evaluation of new and existing hearing aid circuitry—because the inability to understand speech in a noise background is a common complaint of many hearing impaired listeners. Moreover, the need for test items with similar performance-intensity curves is heightened when using adaptive step sizes smaller than the 5 dB step that usually is used clinically. For example, when comparing different hearing aid frequency responses, SRT differences much smaller than 5 dB might be expected (e.g., Van Tasell, Larsen, & Fabry, 1988).

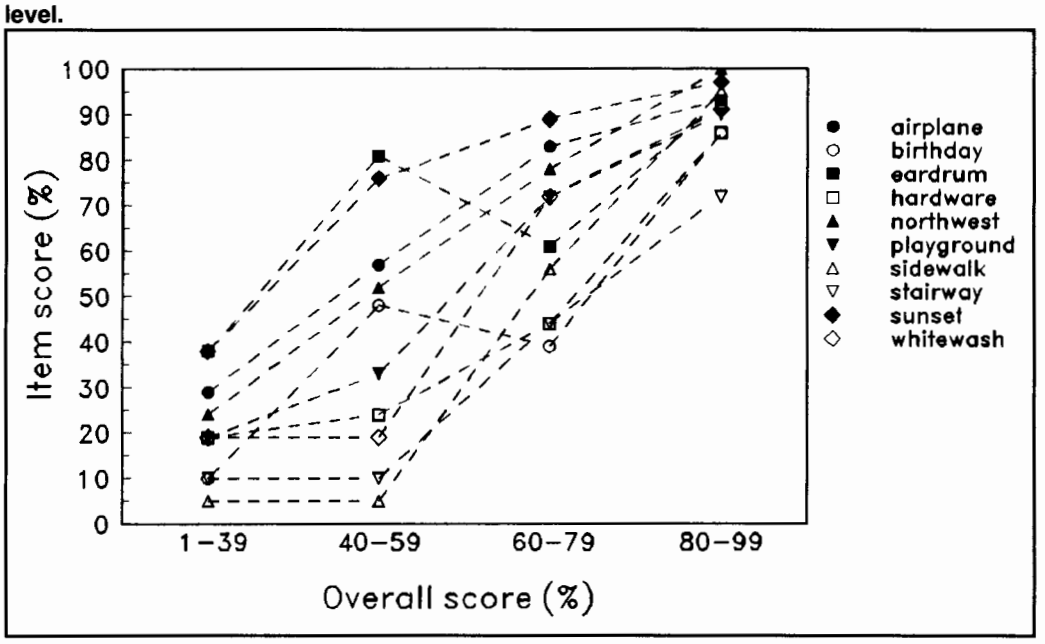
The remainder of this paper will describe how a set of spondees that meet these two criteria for listening in noise were selected and how the adaptive SRT has been applied. In view of their widespread use, easy availability, high face validity, and neutral pronunciation, spondaic words and babble noise recordings from the Auditec catalog were selected for use in this study.

Test Development

Establishment of Homogeneity

Olsen and Matkin (1979) published a summary of six attempts to identify a subset of spondees that were homogeneous with respect to their thresholds of audibility. Each of

Figure 2. Scores obtained for 10 spondaic words as a function of the overall performance level.



the research studies identified different words; only four words (*birthday*, *iceberg*, *northwest*, and *railroad*) were identified as homogeneous by all six studies. This result is not particularly surprising because the relative intelligibility of the words will depend on the test conditions, sound delivery system, and type of masking noise used. However, as a starting point, the 15 words that were identified most frequently by the six researchers as being "satisfactorily" homogeneous (*birthday*, *iceberg*, *northwest*, *railroad*, *playground*, *stairway*, *airplane*, *armchair*, *eardrum*, *farewell*, *hardware*, *mousetrap*, *sidewalk*, *sunset*, and *whitewash*) were selected for further examination in the present work.

Auditec recordings of these 15 spondees were low-pass filtered at 8 Khz and digitized at a sampling rate of 20 Khz using a 16 bit analog-to-digital converter (Ariel DSP-16), edited using CSRE software (Jamieson, Ramji, & Nearey, 1989), and stored for later use. Speech testing was conducted in the sound field in a double-walled IAC sound attenuating booth. Auditec recordings of multi-talker babble were used as the noise source and mixed with the speech for presentation by a speaker located 115 cm in front of the listener. The noise level remained constant at 65 dBA and the level of the speech was adjusted. The speech levels were individually chosen for each listener so a performance-intensity (or psychometric) function was obtained that ranged from perfect performance (operationally defined as better than 95% correct) to unintelligible (defined as poorer than 20% correct). Listeners were tested at seven signal-to-noise ratios (S/N). Eight normal hearing listeners (aged 21-42), who were employees and graduate students in the Department of Communicative Disorders, passed pure tone threshold screening at 20 dB HL (ANSI, 1969).

Spondees were presented from disk via a digital-to-analog converter (DSP-16) and low-pass filtered at 8 kHz, with the presentation level controlled by a programmable attenuator and an Amcron D-75 amplifier. Spondees and taped multi-talker babble were mixed using a Shure M267 mixer before being fed through another Amcron amplifier to the speaker. Listeners were seated in a modified dental chair that had an adjustable neckrest. They were asked to keep their heads steady, resting against the neckrest to minimize the effect of head motion on the signal level. Each of the 15 spondees was played twice at each S/N. Following each stimulus presentation, the response alternatives were displayed on the screen of a computer monitor placed immediately below the speaker. Listeners used a computer mouse to move a pointer that indicated 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. After 30 trials (15 spondees x 2 presentations) were completed, the speech level was changed and a new set of trials commenced.

The results were analyzed to determine how well each spondee was identified *when compared with the overall performance for all 15 spondees*. The overall performance was categorized into four performance levels: 39% correct or less, 40-59% correct, 60-79% correct, and 80% or more correct. At each of these performance levels, the scores were then computed for each individual spondee. Figure 1 displays the result of plotting these scores for individual items as a function of the overall performance level. As can be seen, some spondees were readily identified, even when overall performance was poor, whereas others were difficult to identify,

even when overall performance was very high. For example, identification of the word *railroad* stayed below 50% even when performance on the entire set of items was above 80%.

To develop a set of more homogeneous words, the five words that least reflected the overall performance levels were eliminated, and the evaluation sequence was repeated using only the remaining 10 spondees. The results with this smaller set are plotted in Figure 2. Four more words were eliminated—two (*eardrum* and *sunset*) that had item scores that were better than the overall score and two (*sidewalk* and *stairway*) that had lower item scores than the overall score. This process of elimination was used to maximize the relationship between performance on an individual test item and overall performance levels. The six remaining spondees (*airplane*, *birthday*, *hardware*, *northwest*, *playground*, and *whitewash*) were used in the subsequent adaptive testing procedures.

Adaptive Testing

For adaptive testing, the six spondaic words were used. The level of the noise was held constant, while the level of the speech varied according to the adaptive tracking algorithm. Following each trial, subjects selected the word they heard from a screen display. Each response made by a subject was either correct (i.e., the word presented was selected) or incorrect. Levitt's (1971) adaptive tracking algorithms can be used to track a variety of response levels. For example, for the 50% correct response level, the signal level is decreased after each correct response (to make the task more difficult and move down the psychometric function) and is increased after each incorrect response (to make the task somewhat easier, moving up the psychometric function). By a series of approximations, the method thus converges on the 50% point, at which the probability of moving up the performance-intensity function equals the probability of moving down. A level of 70.7% correct was selected for several reasons. First, with the noise level constant, the algorithm increases speech levels after every incorrect response and decreases the level after two correct responses. The tracking algorithm is computationally simple, but it seems to be relatively difficult for a naive listener to guess what the algorithm is doing. Second, subjects find it more enjoyable to participate when they perform at 70.7% correct rather than at 50% or less because they are less likely to feel that they are merely guessing all the time and become discouraged.

Words were presented initially at a level well above threshold. The level decreased in 3 dB steps after each two correct responses 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 8 reversals were averaged to obtain a threshold estimate and a measure of response variability.

The standard deviation of the 8 reversals was required to be less than or equal to 3 dB, otherwise the adaptive run was repeated. To ensure stability over trials, the final SRT estimate was taken as the mean of the threshold estimates obtained from two consecutive adaptive runs that also did not differ by more than 3 dB. These somewhat arbitrary criteria were judged to be appropriate for research purposes (e.g., Jamieson, Raftery, Cheesman, Ramji, & Keith, 1991). For other potential applications, fewer numbers of reversals may be obtained and/or different criteria for acceptable variability may be adopted.

Test Application

A typical testing session with a hearing aid begins by orienting the listener to the task in a 12-trial practice session during which each of the six spondees is presented twice at a favourable S/N. This practice session permits subjects to become familiar with the six spondees and comfortable with the mouse and monitor response system. This procedure has been used with normal hearing young adults and pre-teen children, and with hearing impaired adults. All listeners have been able to learn to control the mouse during these 12 practice presentations, and the only special difficulty has been that one listener with very low vision was unable to see the words on the screen and could not complete the task.

The SRT procedure described here is highly sensitive, so that relatively small differences in performance will be detected. As an example, to evaluate the test, monaural and binaural SRTs were compared to determine whether the test could detect the anticipated binaural advantage in quiet. A distinct effect was observed for SRT; thresholds for 14 young adult (aged 20-42) normal hearing listeners (pure tone thresholds <20 dB HL, ANSI 1969) were an average of 2.73 dB higher (t=6.45, p <.01) when one ear was plugged and muffed than when both ears were used. As discussed below, this is a listening situation in which higher variability of results might be expected simply on the basis of test item selection procedures.

Because the test is automated, it is straightforward to administer accurately, and data are collected and scored automatically by computer. Because of the pre-specified criteria of acceptance, test-retest reliability is necessarily high—for the SRT to be acceptable, the standard deviation of the measurement cannot exceed 3 dB *within any run* and *two successive runs must fall within 3 dB of each other*.

Notwithstanding these strict criteria of reproducibility, in our preliminary testing, 15 older moderate to moderately-se-

vere sensorineural hearing impaired listeners (aged 57-83 years with a mean of 72 years; educational and employment status unknown) required an average of just 2.23 runs to reach criterion when they were tested in noise (*s.d.* = 0.36 adaptive runs; range = 2-3 runs). In quiet, these listeners had more difficulty, requiring an average of 3.8 runs (*s.d.* = 1.9 runs; range = 2-9 runs). This increased variability when listening in quiet may be a result of the method used to select the test items because the homogeneity of the test items was determined in the multi-talker babble and not in quiet. If tested in quiet, or in other noises, alternative sets of spondaic words would likely be chosen (e.g., ASHA, 1988). In each of their test conditions, younger listeners have required fewer than 3 adaptive runs, on average, to reach criterion. Thus, an estimate of SRT for a given subject in any of the conditions studied to date can normally be obtained in 3-5 minutes. Given this result, the test is suitable for clinical use, even within the very tight time constraints imposed on many clinical environments. Further research is currently being initiated to test the variability of performance in a clinical setting.

Conclusions

A modified version of an adaptive SRT procedure, designed to be used as part of a battery of speech tests, has been developed and found to meet design criteria. One demonstrated application of the procedure is in the evaluation of hearing aids in a laboratory environment (Jamieson et al., 1991). The SRT described here is easy for subjects to learn and do, and it is fast, sensitive, and highly reliable. Subjects with a wide range of age and educational backgrounds have little difficulty with the task after only minimal practice, and they do not show significant changes in performance with repeated testing. Finally, testing and data collection are automated to ensure that the algorithm is precisely implemented and that the tester's time is used to maximum advantage. Because of the computerized response mode, listeners require moderately good corrected vision and manual dexterity.

The advantages of the adaptive SRT approach extend well beyond the laboratory situations in which they have traditionally been applied (Lutman, 1987). As a consequence, this approach could find application in a variety of clinical situations.

Acknowledgements

1. This paper is based upon a talk given at the May 1990 meeting of the Canadian Association of Speech-Language Pathologists and Audiologists. The work reported was supported by Unitron Industries Limited and through grants from the Ontario Ministry of Health and Ontario's University Research Incentive Fund.

2. I am grateful to Drs. Don Jamieson, J.-P. Gagné, and Richard Seewald for their support and encouragement of this work and for helpful discussions. I am also grateful to Shane Moodie for advice and subject referrals and to Ludwig Moser for suggestions regarding the validation of procedures. B. Bentley, J. Keith, J. Beliaeff, E. Baird, V. Masterson, K. Ramji, W. Alsop, and E. Raftery have provided assistance at various stages of this project.

Address all correspondence to: Dr. Margaret F. Cheesman, Hearing Health Care Research Unit, Department of Communicative Disorders, University of Western Ontario, London, N6G 1H1.

References

American National Standards Institute (1969). ANSI S3.6 1969 *Specifications for Audiometers*. American National Standards Institute, Inc., New York.

American Speech-Language-Hearing Association (1988). Guidelines for determining threshold level for speech. *ASHA, 30*, 85-88.

Dirks, D. D., Morgan, D. E., & 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., Ramji, K., & Nearey, T.M. (1989). CSRE: The Canadian Speech Research Environment. *Journal of Speech-Language Pathology & Audiology, 13*, 67.

Jamieson, D.G., Raftery, E., Cheesman, M. F., Ramji, K., & Keith, J. (1991). Simulation and comparison of adaptive hearing aid circuits. International Hearing Aid Conference: Signal Processing, Fitting and Efficacy. Iowa City, IO.

Levitt, H. (1971). Transformed up-down methods in psychoaccoustics. *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 M. Martin (Ed.), *Speech Audiometry*. London: Whurr Publishers.

Martin, M. (1987). *Speech Audiometry*. London: Whurr Publishers.

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

Olsen, W. O., & Matkin, N. D. (1979). Speech Audiometry. In W.F. Rintelmann (Ed.), *Hearing Assessment*. Baltimore: University Park Press.

Thornton, A., & 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. J. (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., & 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.