Development of a Test Environment to Evaluate Performance of Modern Hearing Aid Features

Michael Nilsson*
Robert M. Ghent*
Victor Bray*
Richard Harris†

Abstract
This article describes a new test environment and materials with the potential to measure performance differences among different hearing aid signal processing methods and features. Normative data suggest a linearly predictable increase in difficulty on a speech-in-noise task as the masker changes from random noise to multiple-talker speech, and the number of talkers increases. Data collected with normal listeners revealed no differences across four test sites for the single loudspeaker (Noise-Front) results and some differences across test sites for the multiple loudspeaker results. Room dimension differences among audiometric test enclosures and diffusion (or lack thereof) of the maskers in the vertical and horizontal dimensions of the sound fields appear to account for performance differences for the multiloudspeaker arrays, confirming the need to limit maskers to aperiodic signals in rooms with controlled ceiling height or to establish norms for each test environment such that results obtained in different enclosures can be compared.

Key Words: Diffuse field, directional microphone, hearing aid, HINT, multiple-talker speech, noise reduction, random noise, sound field, speech testing, verification

Abbreviations: 2DD = two-dimensionally diffuse; ANSI = American National Standards Institute; BYU = (test site at) Brigham Young University; CCF = (test site at) Cleveland Clinic Foundation; DAW = digital audio workstation; HIA = Hearing Industries Association; HINT = Hearing in Noise Test; ISO = International Organization for Standardization; ITU = International Telecommunications Union; LTAS = long-term average spectrum; MANOVA = multivariate analysis of variance; MTS = multiple-talker speech; P-I = performance-intensity; rms = root-mean-square; RTS = reception thresholds for sentences; SACD = super audio compact disc; SLC = (test site at) Salt Lake City; SNR = signal-to-noise ratio; SPIN = Speech Perception in Noise test; WAV = digital audio files in Windows Media format; WUSM = (test site at) Washington University School of Medicine

Sumario
Este artículo describe un nuevo ambiente de evaluación y los materiales con el potencial de medir las diferencias de desempeño entre diferentes métodos de procesamiento de la señal para auxiliares auditivos, y sus características. Los datos normativos sugieren un incremento linealmente predecible en la dificultad de una tarea de lenguaje en ruido, conforme el enmascarador cambia de ruido aleatorio a ruido de lenguaje de hablantes múltiples, y conforme aumenta el número de hablantes. La información colectada con sujetos normo-oyentes no reveló diferencias en los cuatro sitios de evaluación.
Laboratory studies of advanced signal processing hearing aids typically evaluate benefit for speech recognition in noise (speech-in-noise) with respect to the factors of audibility, directionality, and noise reduction. Many recent studies have not shown performance differences associated with these signal processing parameters (Elberling et al, 1993; Ludvigsen et al, 1993). This may be due to the lack of true benefit associated with the particular parameter, the lack of sensitivity to parameter performance in the test environment, or a combination of both. The purpose of this experiment was to test the design of a controlled, sensitive environment with the potential for isolating and quantifying benefit attributable to the advanced parameters found in today's hearing aids. Additional motivation comes from the establishment of guidelines by the HIA (Hearing Industries Association) for the substantiation of advertising claims made by hearing aid manufacturers (2003). Such a test system could provide a valid, repeatable, norm-referenced baseline method capable of measuring benefit for a wide range of technologies without unduly biasing results.

Many competing requirements must be considered when designing test materials to assess human performance with hearing aids. Among these are the general experimental design requirements of reliability, sensitivity, and validity. In addition, when quantifying speech recognition in noise (especially with advanced hearing aids), temporal characteristics (used by noise reduction algorithms) and spatial characteristics (used by directional microphone systems) must be controlled or consistently manipulated to allow for parsing of individual contributors to performance.

The test environment described in the present paper was designed around the Hearing in Noise Test (HINT), a well-known, norm-referenced, reliable, and valid test of speech recognition in quiet and in noise (Nilsson et al, 1994). The HINT is a prerecorded test used to measure the RTS (reception thresholds for sentences) in quiet or in noise. The 250 sentences comprising the HINT are of approximately equal length.
(five to seven syllables) and difficulty (first-grade reading level). They were spoken by a male voice actor and digitally recorded. The sentences were equated for difficulty for presentation in quiet or in noise and grouped into phonemically balanced lists of 10 or 20 sentences, providing 25 equivalent 10-sentence lists or 12 equivalent 20-sentence lists (Hanks and Johnson, 1998). The temporal and spectral characteristics of the original masking noise are stationary. The noise spectrum matches the long-term average spectrum (LTAS) of the sentences, thereby maintaining a steep P-I (performance-intensity) function (Studebaker et al, 1987).

The choice of number and length of lists depends on the need for measurement accuracy and testing time. RTSs in noise are measured by fixing the noise level and adaptively varying the sentence level until the sentences can be repeated correctly 50% of the time. RTSs in quiet are obtained using the same adaptive technique with an equivalent noise level of zero. The adaptive procedure avoids nonlinear ceiling and floor effects customarily associated with intelligibility tests given at fixed presentation levels, and optimizes the statistical efficiency of the test (Levitt, 1978). A threshold measurement with a single list usually takes less than two minutes, contributing to the clinical feasibility of the test.

The reliability and sensitivity of the test have been established through use of within-subject repeated measurements of thresholds with different lists (Soli and Nilsson, 1997). The reliability of the HINT compares favorably with a set of Dutch materials (Plomp and Mimpen, 1979) and demonstrates improved reliability over the SPIN (Speech Perception in Noise test) sentences (Gelfand et al, 1988). A P-I function defines the sensitivity of a test: A steeper slope means that a smaller change in presentation level leads to a larger change in speech recognition. The P-I function for the HINT is an 8.9% improvement in speech recognition for each 1 dB change in presentation level in quiet and noise when testing both normal-hearing and hearing-impaired listeners (Soli and Nilsson, 1997). The slope of the P-I function is closer than other sentence materials to previously reported P-I functions that have been measured with word lists (Studebaker and Sherbecoe, 1991). The well-controlled length and simplicity of the sentences limit the contextual effects that account for the steeper functions shown with continuous discourse (Bronkhorst and Plomp, 1989). A tradeoff between reliability and realistic quality of the materials forces the sacrifice of one (the use of continuous discourse and conversational speech) for the other (improved reliability and sensitivity).

Digital noise reduction algorithms in hearing aids monitor auditory signals over time, seeking out steady-state energy within the continuous waveform (Chabries and Bray, 2002). There are two implications for testing speech recognition in noise with noise reduction algorithms. First, the test materials must allow the noise reduction algorithm sufficient time to sample the signal for noise. This requires that any masker must precede the target signal by sufficient time to allow the noise reduction algorithm to engage and stabilize (Nilsson et al, 2000).

Second, noise reduction algorithms act on the more-or-less steady-state elements of a signal, requiring maskers to contain some steady-state energy. There are many steady-state maskers found in the real world (road or traffic noise, fan or motor noise, airplane noise, or crowd noise) as well as many modulated maskers (speech). Specifying one type of masker as representative of all environments is artificial, no matter which masker is chosen; instead, maskers should be chosen for their efficiency or for their physical properties to trigger specific features. The clinically reasonable solution is to sample maskers that extend over a range of conditions to compare and contrast the properties of interest. By defining the hardest condition, or even the easiest, laboratory testing can set boundaries to understand the potential of a system, with real-world benefit occurring somewhere within a defined range. This must be done with well-controlled stimuli in order to increase sensitivity and must be correlated with patients’ experiences in the real world.

Directional microphones used in analog and digital hearing aids require spatial separation between the target signal and the masker in order for benefit to be realized, requiring use of multiple loudspeakers in the test environment. This allows the target signal to be delivered from in front of the subject (0° azimuth, assuming listeners face the person with whom they are talking) with
the competing masker coming from other locations, generally from the sides and back of the subject. Various configurations of noise sources have been reported (Leeuw and Dreschler, 1991; Valente et al, 1995; Voss, 1997; Ricketts and Dhar, 1999; Wouters et al, 1999; Ricketts, 2000; Revit et al, 2002), but while more complex sound fields may better represent the environments encountered in real-world listening situations, complex sound fields are very difficult to replicate in clinical settings.

Limitations must be considered when generalizing laboratory soundfield test conditions to real-world hearing aid performance. Listeners are often exposed to environments where maskers have broad spectra and/or modulated properties (e.g., cocktail party noise). However, many environmental maskers may be steady state and have a limited bandwidth (e.g., motors, fans, road noise). Listeners are often in diffuse or reverberant fields where there is as much noise in the foreground as in the background—few environmental maskers are positioned ideally at or near the polar response nulls of directional microphones—and significant noise from many directions simultaneously may confuse adaptive directional microphone systems, which results in a fixed directional response from the system.

The test environment must reflect its clinical purpose: verification, validation, or both. Verification entails measurement of behavioral performance and quantifiable benefit from the individual hearing aid features. Such a test environment must be repeatable and replicable, well controlled and clinically valid—established in line with accepted norms. For the purposes of validation, such a system should be able to represent as nearly as possible the real-world listening environments in which individual patients might reasonably be expected to communicate, where the patient's unaided performance in the same environment serves as the individual's control. In this case, the test environment need not be replicable or well controlled, and by individual necessity, norms do not exist (Anastasi, 1988).

We propose a test environment that meets the rigors of clinical verification but may also be used to validate hearing aid fittings. In evaluating advanced hearing aids with noise reduction and/or directional microphones, the challenge is to develop a laboratory test that provides careful control over both temporal and spatial variables. This paper describes the development of the materials, equipment, and environment for such a test, and reports on normative performance. A follow-up paper in preparation describes unaided and aided performance in the same test environment with hearing-impaired listeners.

The identical test configuration was installed at four sites. Normative data were collected at all sites to evaluate the equivalency of the test configuration and to allow comparisons of data collected from samples of different normal-hearing populations. To accommodate the operational requirements of the noise reduction and directional features of modern hearing aids, the HINT masker was modified for noise onset time, diffusion, and multisignal correlation. A new set of maskers was created using multiple-talker speech. Details of these modifications are found in the “Methods” section below.

**METHODS**

**Test Materials**

**Hearing-in-Noise Test**

The HINT masking noise is shaped to match the LTAS of the male talker in an attempt to reduce auditory cues elicited by spectral differences between the target signal and the masker (Nilsson et al, 1994). Several modifications were made to the HINT masking noise for the present study.

The first modification was to move the start of the noise forward in time relative to the start of each sentence by increasing the time interval between the onset of the noise and the onset of the sentences from 0.5 seconds to approximately 5.0 seconds. (The original intention of this modification was to allow time for single-microphone noise reduction algorithms to engage, but effects on normal performance were also observed and will be discussed later.) Evaluation of single-microphone adaptive noise reduction algorithms can be accomplished with a single loudspeaker in this design.

The second and third modifications to
the HINT noise involved the creation of diffuse, correlated and diffuse, uncorrelated masker tracks to be reproduced simultaneously from four different loudspeakers arranged in a soundfield array. A single noise signal delivered through all four loudspeakers simultaneously creates a correlated noise field. Four separate recording tracks of noise, each having a different starting point when sampled from the original HINT noise recording and delivered through the individual loudspeakers surrounding each listener, create an uncorrelated noise field.

Binaural presentation of correlated signals in a sound field or under headphones creates a fused auditory image in the center of a listener’s head, whereas uncorrelated maskers are perceived as originating outside a listener’s head. These phenomena (well-known and described in myriad other references) were confirmed in the present test environment using the HINT masking noise. Data were collected with normal listeners using the correlated noise presented from multiple loudspeakers to compare its effect on speech-in-noise performance to that of uncorrelated noise in the same test environment.

A new set of maskers was created by combining recordings of uncorrelated multiple-talker speech, with four, eight, twelve, and sixteen individual talkers. This masking condition adds the variable of temporal modulation across different numbers of talkers. Details on the creation of these new maskers follow.

**Multiple-Talker Speech**

Recordings of eight young adult male and eight young adult female talkers were randomly selected from a sample of 50 individual digital recordings of each gender (100 total recordings) reading the “Television Passage” (Cox and Moore, 1988). The original recordings had been made in an anechoic room with a free-field microphone positioned 30 cm away from each talker (Layton, 1991).

All of the original recordings were transferred onto a computer hard disk and converted to Windows Media format digital audio (WAV) files at 16-bit resolution and a sampling rate of 44.1 kHz using Sound Forge 4.5 (Sonic Foundry). The files remained in the digital domain, at this bit resolution and sample rate, throughout editing, mixing, file transfer, and final reproduction for subject testing.

Editing was accomplished using Cool Edit Pro 1.2 (Syntrillium Software). Each passage was edited to remove pauses, breaths, and discontinuities (>75 msec) in order to create continuous, intelligible, running speech. Temporal gaps due to the morphemic, phonemic, and prosodic features of speech were retained.

Since the original masking noise of the HINT was filtered to match the LTAS of the male talker for the target sentences, the recordings of the individual talkers were likewise filtered to approximate the LTAS of the HINT noise. The result of this filtering was to remove from the recording of each talker any energy not present in the HINT noise spectrum. However, energy present in the HINT noise spectrum (predominantly low-frequency) was not added to the spectra of the talkers where missing. This produced small changes to the sibilant energy for each talker but did not alter the formant structure, such that individual talkers could still be uniquely identified. As the talkers were summed to create the multiple-talker files, the resulting waveforms and spectra became increasingly like those of the HINT noise (Figures 1 and 2).

The recorded passage was uncorrelated among the 16 talkers by setting the starting point for each talker at a different position within the passage. The individual recordings were trimmed to the approximate length of time required to deliver two sequential lists of ten HINT sentences, including the increased noise onset time and an appropriate interstimulus interval for subject response (about 4.2 total minutes for each 20-sentence list). The recordings were then mixed to achieve the desired number of multiple talkers per test condition, with an equal average rms (root-mean-square) power contribution from each talker. Each test condition was equated to within 0.1 dB of the average rms power of the HINT noise. The mix of talkers for each test condition was gender-balanced, and there was no talker redundancy in any of the multiple-talker maskers.

Finally, each recording was synchronized with the presentation of the HINT sentences with silence inserted at the interstimulus intervals. The synchronized tests were saved
Figure 1. Spectra of maskers: (a) a single talker for reference; (b) four-talker MTS; (c) eight-talker MTS; (d) twelve-talker MTS; (e) sixteen-talker MTS; (f) HINT noise. All of the MTS maskers were filtered to match the LTAS of the HINT noise and obtained using a 1024-line FFT with a Blackman-Harris window.

Figure 2. Time-waveforms of maskers: (a) a single talker for reference; (b) four-talker MTS; (c) eight-talker MTS; (d) twelve-talker MTS; (e) sixteen-talker MTS; (f) HINT noise. All waveforms are ten-second samples and equated for average rms power.
as WAV files on recordable compact disc and uploaded to a Yamaha AW4416 DAW (digital audio workstation) for final mixing, routing, and playback in the sound field.

**Test Equipment**

**Multitrack Recording and Playback**

The Yamaha AW4416 DAW is a 16-track, hard disk-based digital mixer and recorder. Each HINT list in the present study consisted of two sets of HINT sentences (a total of 20 sentences per list). Signal routing and final mixing levels for playback were precalibrated as established in the sound field by the fader (volume control slider) positions for each test condition (quiet, HINT noise from the front, HINT noise from the loudspeaker array, and four-, eight-, twelve-, and sixteen-talker speech from the loudspeaker array). Mixer level and signal routing settings were saved to the hard disk along with each HINT list. Playback of a particular test condition required selecting a HINT list from a menu, then recalling the settings from memory.

**Loudspeaker Array**

Six discrete outputs from the DAW were used to deliver signals to the sound field. Two outputs were routed to channels one and two of the audiometer for control of signals (target and masker) to be presented from the center/front loudspeaker of the soundfield array. The left soundfield output of the audiometer was routed to one channel of a five-channel power amplifier. Four other assignable outputs of the DAW were routed to the four remaining channels of the power amplifier.

The outputs of the power amplifier channels were used to drive the five loudspeakers of the soundfield array, arranged according to Figure 3. All five of the loudspeakers (RCA PRO-X44AV) were positioned on stands 1 m away from the center of the listening position per ANSI (American National Standards Institute) requirements (1996) and ISO (International Organization for Standardization) requirements (1992). Speaker height was selected to match average adult head level when seated (1.15 m). The four loudspeakers...
used for reproducing the maskers were equidistant from each other and set at 45°, 135°, 225°, and 315° azimuth.

The system was calibrated such that the HINT masking noise was reproduced at 65 dB(A) at the center of the listening position under any test condition (any masker from the front loudspeaker or any masker from the soundfield array). When the soundfield array was used to reproduce the masking condition (both HINT noise and multiple-talker speech [MTS]), the mixer settings ensured equal contribution from each loudspeaker.

While the sound field met the ANSI (1996) and ISO (1992) requirements for a quasi-free sound field, this definition does not adequately describe the behavior of the test environment, nor is the term “quasi-free sound field” an accurate label for what is described in the two standards. A free sound field, by definition, has no boundaries. Typical audiometric test environments are enclosed volumes with boundaries that exhibit absorption and reflection to varying degrees at different frequencies. By placing uncorrelated sound sources around the azimuth of the listening position, the effect of diffusion can be created within a subspace of the enclosed volume (see Appendix by Ghent, 2004, in this issue), but only in the two dimensions represented azimuthally. Therefore, the sound field thus created may be defined, with limitations, as two-dimensionally diffuse, or 2DD.

The design rationale for using five loudspeakers in the array described in Figure 3 was based on several factors. Quantifiable verification of speech-in-noise performance requires controlled laboratory conditions, specifically a diffuse sound field at the listening position. Having four evenly spaced loudspeakers for delivering the masking signal satisfies this requirement, creating a stable 2DD environment. Should it be desirable to validate individual hearing aid fittings, however, where simulation of real-world listening environments is desired, the same five loudspeakers can be spaced for reproduction of surround-encoded audio according to the International Telecommunications Union Recommendation ITU-R BS.775-1 (1994). Audio production techniques for creating realistic simulated audio environments using the recommended five-loudspeaker array, without the need for accompanying motion picture visual reinforcement, have been described (Bosun, 2001), and the five-loudspeaker array thus arranged has been shown to create a simulated listening environment that is nearly equivalent psychoacoustically to a 12-loudspeaker reference array (Hiyama et al, 2002), while being much less costly and complex.

**Test Environments**

The test setup described was replicated at four sites: the Hearing Aid Research Laboratory at Sonic Innovations in Salt Lake City, Utah (SLC), the Department of Otolaryngology at Washington University School of Medicine in St. Louis, Missouri (WUSM), the Audiology Research Laboratory at the Cleveland Clinic Foundation in Cleveland, Ohio (CCF), and the Department of Audiology and Speech-Language Pathology at Brigham Young University in Provo, Utah (BYU). The audiometric test room at each site had to be large enough to accommodate the five-loudspeaker soundfield array shown in Figure 3 (a minimum inside dimension of 2 m by 2 m). The listening position was located at the center of each audiometric test room. The audiometric test rooms complied with ANSI requirements for permissible ambient noise, ears uncovered (ANSI, 1991).

Replication of the equipment setup at the various test sites was accomplished using identical equipment and bit-for-bit copies of the digitally recorded test materials, including the precalibrated mixer settings. Calibration at the listening position of each of the alternate test sites was verified to be within 0.5 dB of the original test site. The spectra of the HINT maskers were measured in the Noise-Front and uncorrelated 2DD soundfield conditions at the listening position of three of the four test sites. The 1/3-octave spectra were obtained using a Brüel & Kjaer 2260 Investigator connected to a 2.54 cm microphone and are shown in Figure 4. These measurements were undertaken to ensure that the HINT noise spectrum (including the peaks at 1 kHz, 3 kHz, and 6 kHz in the FFT-derived spectrum in Figure 2 and the 1/3-octave spectrum in Figure 4) was maintained between the Noise-Front and 2DD soundfield conditions as well as across test sites. The consistency between 500 Hz and 8 kHz across sites was high enough that collection of spectral data at the fourth site was deemed
unnecessary and cannot be collected as the booth has since been disassembled and no longer exists in its original configuration.

**Test Procedure**

Modified hyper-latin-squares were used to randomize the starting order and presentation order of the HINT lists, and the starting order and presentation order of the test conditions. The prescribed adaptive presentation and scoring procedure for the HINT was followed, as described in the HINT instructions. The presentation level of the HINT sentences was varied by adjusting the HL attenuator on Channel 1 of the audiometer.

**RESULTS AND DISCUSSION**

Soundfield test results have been difficult to compare across test facilities because of interactions with the sound field itself; different room sizes and differences in reflecting surfaces within an audiometric test room can change base performance (Ghent, 2004, in this issue). Therefore, the validity of the present set of materials was established across several levels. First, questions concerning noise onset time and correlation of noise in the multiple loudspeaker soundfield conditions were resolved to determine the test conditions to be used in the multisite testing. Next, the difference between the Noise-Front and 2DD soundfield noise conditions was established across the four test setups. Finally, the change in performance with the different multiple-talker speech maskers was evaluated.

**HINT Noise Using Two Noise Onset Times**

Normative data were collected on ten normal-hearing listeners (three male and seven female with pure-tone thresholds less than 20 dB HL between 250 and 6000 Hz, native speakers of English, age range from 24 to 47 years with a mean age of 30.7 years) at the SLC site. Subjects were placed at the listening position, and target sentences were presented from the front loudspeaker while uncorrelated HINT noise was delivered from the four surrounding loudspeakers. Adaptive
thresholds were measured using the 20-sentence lists with both 0.5 second noise onset time and 5.0 second noise onset time.

An ANOVA found a significant difference in thresholds for the two noise onset times \[F(1, 9) = 38.55, p < .01\]. The longer noise onset time produced lower thresholds (-8.16 dB SNR [signal-to-noise ratio]) than the short onset time (-5.58 dB SNR). It is possible that allocation of attention is better accomplished with sufficient time to prepare for listening in this difficult condition, thereby improving thresholds by over 2.5 dB. All multisite data were therefore collected with the long noise onset time, which benefits absolute performance but also allows evaluation of signal processing algorithms that require processing time to stabilize.

The longer noise onset time can also be considered as an accommodation for the limitations of the HINT procedure. Background noise is turned off between sentences in order to hear the subject’s response and determine the presentation level of the next sentence. Real-world situations do not typically include interrupted maskers that occur only when a speech target is present. The longer onset time re-establishes the more typical listening situation of a continuous masker. The clinical implication that should not be overlooked is that using the short onset time in the unaided condition (for clinical expediency) and the long onset time in the aided condition will artificially inflate any observed benefit, or show benefit where none exists.

**Correlated versus Uncorrelated Noise Sources**

Speech-in-noise performance with a single noise signal delivered through all four loudspeakers simultaneously (a correlated noise field condition) was compared to that of four separate tracks of noise, each having a different starting phase, delivered through the individual loudspeakers surrounding each listener (an uncorrelated noise field condition). The same 10 subjects were tested with a five-second noise onset time in both conditions. An ANOVA found significant differences between the correlated and uncorrelated noise field conditions \[F(1, 9) = 6.28, p < .05\], with lower thresholds in the correlated noise condition (-8.16 dB SNR) than in the uncorrelated noise condition (-7.27 dB SNR). Since one goal of the system is to provide laboratory control over conditions that are representative of those that may be encountered in the real world, the use of the uncorrelated noise in the sound field is suggested. It is important to note that differences in the time alignment of even a random noise can cause changes in performance (cf. masking level differences), suggesting that correlated and uncorrelated soundfield conditions should not be directly compared.

**Testing at Multiple Sites**

Thresholds were measured on normal-hearing subjects in one single-loudspeaker condition, and five multiple-loudspeaker conditions at the four test sites (SLC, WUSM, CCF, and BYU). Of interest was the ability of the replicated equipment setup (including digital recorder, speakers, recordings, and room layout) to produce comparable performance in a range of listening conditions. At the same time, differences in absolute performance with various maskers could also be evaluated. All conditions used the five-second noise onset time, and the multiple-loudspeaker soundfield conditions all used uncorrelated signals.

An ANOVA was used to evaluate the single-loudspeaker condition with HINT speech and noise presented from the front loudspeaker. No significant difference between sites was found \[F(3, 46) = 0.6, p = .62\]. The mean RTS for the normal-hearing subjects across sites was -2.9 dB SNR.

A MANOVA (multivariate analysis of variance) was used to compare the four test sites and the five different multiple-loudspeaker conditions (uncorrelated HINT noise presented from four loudspeakers, and one, two, three, or four individual talkers played out of each of the four soundfield loudspeakers creating four-, eight-, twelve-, or sixteen-talker speech maskers). A main effect of test site \[F(3, 46) = 9.1, p < .01\] was found, with a post-hoc analysis revealing that SLC (-5.1 dB SNR) was different from the other three sites (BYU = -3.5 dB SNR; WSM = -3.5 dB SNR; CCF = -3.8 dB SNR). A main effect of masker type was found \[F(4, 184) = 71.0, p < .01\], with post-hoc analyses showing the five maskers produced significantly different RTSs (2DD noise = -5.4 dB SNR; four talkers = -4.7 dB SNR; eight
talkers = -4.0 dB SNR; twelve talkers = -3.3 dB SNR; sixteen talkers = -2.1 dB SNR). No significant interaction was found \([F(12, 184) = 1.08, p = .38]\).

The respective statistical analyses reduce to several succinct observations: (1) No site effect was shown in the single-loudspeaker condition; (2) A site effect was found in the multiple-loudspeaker conditions; (3) RTS values were different for the front versus 2DD noise; (4) variation in the number of talkers produced different RTS values; (5) RTS values were lower for 2DD noise than for all multiple-talker masker conditions.

The consistency of RTS scores across sites in the single-loudspeaker noise condition supports the ability of this test configuration to be accurately replicated and calibrated at multiple sites. This also suggests that the subject pools were equivalent, and differences found in other conditions should not be attributed to population sampling errors.

The major difference among the test sites that could affect the multiple-loudspeaker conditions is the variation in room size, specifically dimensions of the respective audiometric test enclosures. The audiometric test suite at SLC has a volume of 22.79 m³; the test suite at BYU has a volume of 17.52 m³; and the enclosures at WUSM and CCF have volumes of 14.05 m³ and 13.78 m³, respectively. The effect of enclosure volume on HINT RTSs in noise is shown in Figure 5.

Performance with the multiple-talker maskers appears consistent and predictable, and step-wise regressions were used to determine which room dimensions were significant predictors of performance. Factors included in the analyses were room height, depth, and width; largest and smallest distance between any loudspeaker and the walls of the test enclosure; and distance from the front loudspeaker to the closest wall. Backwards regression after a full loading of the dependent variables yields significant predictive relationships between either room depth or width and the multiple-talker conditions \([F(1, 48) = 9.6 to 12.9, p < .01 and r^2 = 0.2 for all analyses]\). Backwards regression after a full loading of the same dependent variables listed above yields a significant predictive relationship between 2DD noise and ceiling height \([F(1, 48) = 25.3, p < .01, r^2 = 0.3]\).

**Figure 5.** Performance (RTSs in noise) with various maskers presented from the loudspeaker array, as a function of the internal volume of the audiometric test enclosures. The respective internal volumes and ceiling heights of the test enclosures at each site are as follows. SLC: 22.79 m³ and 2.44 m; BYU: 17.52 m³ and 1.98 m; WUS: 14.05 m³ and 1.98 m; CCF: 13.78 m³ and 1.98 m. BBN = broadband noise.
Therefore, variability in the RTS scores for the multiple-talker conditions is best predicted by the length-width dimensions of the various test enclosures. The variability in RTS for the 2DD noise is dominated by ceiling height (2.44 m at SLC versus 1.98 m at all of the other sites) where there is no control of diffusion. Unlike the multiple-talker maskers, the lack of periodicity in the random noise means true diffusion (random incidence) is achieved in the two horizontal dimensions controlled by the uncorrelated signal delivery. If the third (vertical) dimension were to be controlled by placement of loudspeakers delivering uncorrelated noise above and below the listening position, the room dimensions would no longer significantly predict RTS in random noise. There would be little change, however, in the relationships for the multiple-talker conditions because their periodic structure would continue to interact with the overall volume of the enclosure (Ghent, 2004, in this issue). Therefore, the limitation of the enclosed 2DD sound field is that diffusion can only be achieved in the controlled dimensions with uncorrelated random energy.

Regarding the observation that the RTS varies between Noise-Front (-2.9 dB SNR) and 2DD noise (-5.4 dB SNR), previous research has long established improved performance in noise due to spatial separation of the target signal and masker for both broadband noise (Licklider, 1948; Dubno, Ahlstrom, 2002) and multiple-talker speech (Gelfand et al, 1988). The present data simply confirm this fact.

Testing with different numbers of talkers in a multiple-talker speech masker was done to test the theory that with fewer talkers in the background, a listener may be able to take advantage of temporal gaps in the masker to extract information from the target signal and complete the message through auditory closure. Recent work by Dubno, Horwitz (2002, 2003) supports this theory, although interrupted noise was used as the masker instead of speech. The theory included the notion that increasing the number of talkers would increase the difficulty of the task as the masker became more gapless. The present experiment supports this notion.

Regarding the observation that RTS was higher for all multiple-talker masker conditions than for the 2DD noise condition, it is most likely other acoustical differences between the maskers, namely the smaller peak-to-power ratio and random nature of the noise versus the similar peak-to-power ratio and periodicity of the multiple-talker maskers to the target that account for the differences, or contextual differences, between the maskers.

The random noise, with its more consistent distribution over time, will have a smaller peak-to-power ratio (a smaller range from the highest intensity peak to the lowest) than is seen in any of the multiple-talker speech maskers (compare the time-waveforms in Figure 2). At some point the peak intensity may be determining the amount of masking provided, which is typically seen with very short speech samples such as individual words (Zhang and Zeng, 1997), especially in the higher frequencies. Therefore, listeners can take advantage of some gaps (which explains why better performance is found with fewer talkers in the masker), but because the alignment of the target and masker is random, the effective masking compared to random noise is greater because of the wider range of levels that are masked.

In 1969, Carhart et al introduced the term “perceptual masking” to explain their finding that a speech-in-noise task increased in complexity as the background (masker) required the listener to disentangle the target from increasing numbers of competing signals comprised of meaningful speech. Data from the present experiment would seem to support this model. However, an experiment by Lewis et al (1988), while confirming again that a speech-based masker was more efficient than broadband noise, did not find support for the perceptual masking model of Carhart et al (1969). Lewis et al (1988), while offering many possible explanations, could not account for their observed differences; however, they did observe that the P-I functions of the noise and babble results differed by a constant (2.3 dB) and that the slopes of the functions were nearly identical as is shown in Figure 5 with the current data.

**CONCLUSIONS**

Modifications to the HINT have been developed that control the temporal and spatial characteristics that are used by features available in modern hearing aids.
Performance characteristics with these devices historically have been difficult to assess because traditional test environments and materials do not adequately address their capabilities and limitations. The new test environment and materials described in this paper increase the noise onset interval, allowing temporal processes sufficient time to take effect, while the introduction of new masking conditions presented from a loudspeaker array allow evaluation of hearing aid features that rely on spatial factors. The new test environment and materials also meet or exceed the requirements set forth in the HIA guidelines (2003).

The materials and procedures described herein were replicated at four laboratories. The Noise-Front condition does not exhibit variation with respect to dimensions of the audiometric enclosure. A consistent shift in performance as a function of the enclosure width and depth was observed for the materials masked by multiple-talker speech. A shift in performance that is dominated by ceiling height was demonstrated with the uncorrelated noise presented in the 2DD sound field. Although both the 2DD and the multiple-talker maskers met the ISO and ANSI definition of a “quasi-free sound field,” we found that room size affected the results, which was unexpected.

With variation in room size, the aperiodic noise generated diffusion in two dimensions (i.e., the results are independent of the room width and depth) but failed to generate diffusion in the third dimension (i.e., the results interacted with room height). This interaction will be a natural consequence whenever a horizontal loudspeaker array is placed at ear level, rather than at varying heights around the enclosure. This suggests that additional uncorrelated noise introducing diffusion in the vertical dimension would eliminate any interaction between RTS and room size. Additionally, standard audiometric test rooms with ceiling heights comparable to the the CCF, WUMC, and BYU rooms (1.98 m) should generate normative data comparable to the current study. The regression equation found can also be used to correct for nonstandard ceiling heights.

On the other hand, with periodic signals, diffusion cannot be generated to the same extent as random noise. The elimination of interactions between RTS and horizontal room dimensions was not attained with these speech materials. Therefore, the cost of using speech as an experimental masker is the loss of control (higher variability) associated with room dimensions, which is not addressed in the current definition of the “quasi-free sound field.” The regression equation predicts that performance will shift by 1 dB for each 6.3 m³ change in volume, and can also be used to adjust the current data to the size of any particular audiometric test room.

Despite the limitation associated with room size described above, the test configuration still addresses our other development criteria (control of spatial separation, noise onset time, and temporal characteristics). This limitation does not prevent the evaluation of absolute performance, which can be attained by comparison of site-specific data to normative data for that site. Analysis of combined group data is possible by using site-specific controls (subjects running as their own controls) and can be verified by the lack of significant interaction between site and the variable of interest. The current norms can be applied to other audiometric test rooms with the caveats that a minimum size is needed for the speaker array, and the appropriate ceiling height or room volume adjustment is applied (based upon the regression equation).

As of this writing, the test environment and materials described in this paper have been used with hearing-impaired listeners to evaluate digital noise reduction, directional microphones, and multichannel signal processing algorithms in hearing aids. These multisite clinical trial results will help us understand how temporal and spatial characteristics of the masking noise affect aided speech recognition with advanced signal processing.

By standardizing the test setup on one of the commonly available 5.1-channel surround-sound configurations, we foresee this methodology being cost-effectively utilized in various clinical settings. Encoding for 5.1-channel sound reproduction is now available on super audio compact discs (SACDs) and audio DVDs, as well as on the audio tracks of conventional DVDs. It is possible to use the test system described in the present study to develop standardized test materials that can then be distributed on the commercially available digital audio media listed above. For general clinical utility, less costly home theater equipment designed
for playback of 5.1-channel audio can be dropped in place of the multitrack DAW used for test material development, mixing, and encoding. Additionally, on the DVD format many alternate maskers can be included, as well as tests in multiple languages. Simulations of real-world listening environments can be menu-accessed and enhanced with the addition of motion picture visual reinforcement for use as a validation tool.

Acknowledgments. The authors wish to thank Jane Enrietto and Suzanne Kornhass for data collection at WUSM and CCF, respectively; Jeannette Johnson for her contributions to protocol development; Harvey Dillon for his suggestions regarding technical issues; and Patrick Murphy, Michael Valente, and David Fabry for their reviews of and comments on the experimental design. Special thanks to Roxanne Olsen and Rebecca Nilsson for their help in preparing data collection materials, and to Jeff Tackett for helping us get over the learning curve on the Yamaha equipment faster than we would have otherwise.

REFERENCES


