Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise

Michael Nilsson, Sigfrid D. Soli, and Jean A. Sullivan
House Ear Institute, 2100 West Third Street, Los Angeles, California 90057

(Received 16 June 1993; revised 8 September 1993; accepted 25 October 1993)

A large set of sentence materials, chosen for their uniformity in length and representation of natural speech, has been developed for the measurement of sentence speech reception thresholds (sSRTs). The mean-squared level of each digitally recorded sentence was adjusted to equate intelligibility when presented in spectrally matched noise to normal-hearing listeners. These materials were cast into 25 phonemically balanced lists of ten sentences for adaptive measurement of sentence sSRTs. The 95% confidence interval for these measurements is ±2.98 dB for sSRTs in quiet and ±2.41 dB for sSRTs in noise, as defined by the variability of repeated measures with different lists. Average sSRTs in quiet were 23.91 dB(A). Average sSRTs in 72 dB(A) noise were 69.08 dB(A), or −2.92 dB signal/noise ratio. Low-pass filtering increased sSRTs slightly in quiet and noise as the 4- and 8-kHz octave bands were eliminated. Much larger increases in SRT occurred when the 2-kHz octave band was eliminated, and bandwidth dropped below 2.5 kHz. Reliability was not degraded substantially until bandwidth dropped below 2.5 kHz. The statistical reliability and efficiency of the test suit it to practical applications in which measures of speech intelligibility are required.

PACS numbers: 43.71.Gv

INTRODUCTION

Traditional assessments of hearing loss based on pure-tone thresholds may not adequately measure the function of the auditory system with wideband signals, nor do these assessments accurately predict speech intelligibility in noisy environments (e.g., American Academy of Otolaryngology committee on Hearing and Equilibrium, 1979). The current research describes a reliable and efficient method for direct assessment of speech intelligibility from speech reception thresholds measured in quiet and in noise.

I. PREVIOUS RESEARCH

Several existing tests can be used to assess speech communication in noise, e.g., the Connected Speech Test, Cox *et al.*, 1987; the City University of New York topic-related sentences, Boothroyd *et al.*, 1985; and the Speech Perception in Noise Test, Kalikow *et al.*, 1977. These tests have been designed to measure percent intelligibility at fixed speech and/or noise levels. The tests produce reliable estimates of performance, but percent intelligibility measures are inherently limited by floor and ceiling effects. These limitations are most evident when the test is administered under conditions where measured intelligibility deviates from 50% correct. An alternative to percent intelligibility measures is the speech reception threshold (SRT) which is not subject to floor and ceiling effects. The SRT is defined as the presentation level necessary for a listener to recognize the speech materials correctly a specified percent of the time, usually 50%.

The technique for measuring SRTs is derived from adaptive testing where the presentation level of the stimulus is increased or decreased by a fixed amount, depending upon the listener's ability to repeat the material correctly. An adaptive SRT method, as noted by Levitt (1978, p. 287), offers effective placement of presentation levels within the region of interest with "concomitant improvements in the efficiency and accuracy of the estimation." Over a sequence of trials, the level of a subsequent stimulus is increased when the response to the current stimulus is incorrect, and, likewise, the level of a subsequent stimulus is decreased when the current response is correct. In this way, the presentation level will approach the listener's SRT. The SRT is estimated by averaging the presentation levels in the latter part of the sequence. Implicit to the measurement of adaptive thresholds with speech materials, however, is the requirement that the materials used in each trial of an adaptive sequence be of equal, known difficulty. Since speech materials become less difficult as they are repeated or reused, this requirement also means that the speech materials must be different for each trial.

SRTs are often measured in the presence of noise (either shaped random noise or multitalker babble). The speech-to-noise (S/N) ratio at threshold can be used in these instances to compare SRTs measured at different speech or noise presentation levels. Dirks *et al.* (1982) measured percent intelligibility and SRTs in 12-talker babble using 36 spondees. Their normal-hearing listeners showed good agreement between intelligibility, as measured with a performance-intensity function, and the SRT, as measured with an adaptive procedure. They concluded that the SRT can provide the same information as the lengthier testing procedures required for performance-intensity functions.

Spondee word lists have frequently been used in clinical settings to measure SRTs because they are faster and easier to administer than sentence materials (e.g., Carhart, 1946). But in order to maximize the reliability of such measures, the individual spondees must be equated for difficulty. Hirsh *et al.* (1952) have shown that the spondees of the W1 and W2 lists are not of equivalent difficulty, and previous attempts to find equivalent subsets of spondees have not come to a consistent agreement. Thresholds are reliable as long as they are estimated using the entire set of 36 spondees from the CID W-1 and W-2 lists. This procedure may not be practical because of the time required for test administration and training. The listener must also be familiar with the set of spondees to minimize learning effects between multiple test conditions.

Spondees are also less representative of "natural" language communication than sentences, since spondees spoken as isolated utterances or in carrier phrases may not represent the normal spectral weighting, level fluctuations, intonations, pauses, etc., associated with conversational speech. Moreover, the limited number of spondees together with the risks of familiarization and learning effects associated with randomization and reuse of the same items prevents measurement and comparison of performance in multiple experimental or clinical conditions. Finally, if measurements are to be taken with listeners using auditory prostheses, the duration of a spondee utterance may not be sufficient to engage dynamic processing characteristics such as compressors found in many contemporary auditory devices. These limitations underscore the need for sentence-length test materials that can be used to measure SRTs.

Sentence-length materials have been shown to be efficient by Hagerman (1982, 1984), who created a set of Swedish sentences for use in an SRT task. Plomp and Mimpen (1979) have created a set of Dutch sentence materials that have been widely used, while Laurence *et al.* (1983) have adapted a British sentence test for measuring SRTs. Hagerman found sentence SRTs (sSRTs) in noise were highly repeatable, with the standard deviation of the error between repeated measures to be 0.44 dB. His materials were created with the same words reorganized to create alternate sentences; however, listeners commented that some sentences were nonsensical and unnatural. In addition, significant learning effects were observed because of the repeated use of the same words. As an alternative, Plomp and Mimpen, as well as Laurence *et al.*, have obtained reliable results, with standard deviations of error scores of 0.9 and 1.4 dB, respectively, by using carefully selected natural sentences from a large corpus of materials without repetitions.

SRTs have also been measured with American English sentence materials. Dubno *et al.* (1984) used the Speech Perception in Noise (SPIN) sentences (Kalikow *et al.*, 1977; Bilger *et al.*, 1984) and spondees to measure SRTs in quiet and in 12-talker babble. SRTs were measured by fixing the speech level and adapting the noise level based on performance for the final word in the SPIN sentences. Reliability of the sSRTs was comparable to Plomp and Mimpen's (1979) findings, but the limited number of SPIN sentence lists and the use of single word scoring for each sentence reduces the efficiency of the SPIN test for SRT measures. Gelfand *et al.* (1988) used only the high predictability sentences of the SPIN, but scored the entire sentence in an SRT task. Their approach yielded a signed average of difference scores less than 1 dB, though only ten lists are available with this approach, limiting the number of conditions that can be tested without repetition of lists.

The foregoing limitations in American English sentence materials have led us to develop a set of sentence materials specifically for use in SRT measurements. We began with the Bamford–Kowal–Bench (BKB) sentences (Bench and Bamford, 1979), which are a large set of short sentences designed for use with British children. The sentences incorporate common nouns and verbs found in transcriptions of British children's speech, and are designed to be scored based on recognition of key words. The size of the set, allowing the development of a large number of alternative lists, as well as its simplicity and brevity, made this test a desirable starting point.

The BKB sentences contain a number of British idioms and usages that detract from their naturalness to American English listeners. The sentences also vary in length over a range that may influence the ability of listeners to remember and repeat the entire sentence correctly, which could introduce memory effects into the measurements. The remainder of this paper describes the development of an sSRT test for use with American English listeners based on the BKB materials. This test, entitled the hearing in noise test (HINT), is composed of 25 equivalent lists of ten sentences that have been normed for naturalness, difficulty, and reliability.

II. CREATION OF TEST MATERIALS

A. Development of sentence materials

The BKB sentence materials have previously been used for SRT measurements (Macleod and Summerfield, 1987, 1990; Laurence *et al.*, 1983), but no specific modifications or norming has been done to equate sentence difficulty, as required in an adaptive SRT task. The sentences also required revisions to eliminate British idioms and usages, and to obtain uniform sentence lengths of six or seven syllables.

1. Procedure and results

The 336 BKB sentences were revised to remove British idioms and to equate their lengths. After these revisions, ten native speakers of American English evaluated the sentences for naturalness on a seven point scale (7 ="natural," 1="artificial"). The subjects were asked to provide suggestions for changes that would make the unnatural sentences more natural sounding. Any sentences that did not receive a mean rating of at least six were revised using the subjects' suggestions. Verb tense was altered to equate the number of sentences written in the past and present tense. All revised sentences were rated again by another set of six subjects. The second set of revisions was adequate to yield mean naturalness ratings of six or above for the revised sentences.

Recordings were made of the revised materials using a male professional voice actor. The sentences were sampled directly to disk at 20161 Hz using an Ariel digital signal processing board with a TMS320 processor and 16-bit A/D and D/A converters. Dual, cascaded antialiasing filters with a roll-off of 96 dB/oct were set to 8 kHz. The voice actor was instructed to maintain clarity, pace, and effort. Recordings were made in a double-walled sound room with acoustic foam on the walls and ceiling. A 1-in. Bruel and Kjaer (B&K) microphone (Type 4144) was placed perpendicular to the talker at a distance of 1 m. A B&K frequency analyzer (Type 2121) modified to include an 80-Hz high-pass filter was used to amplify the microphone signals. Average signal levels at the microphone were maintained at about 70 dB SPL. Signal levels were monitored with an oscilloscope throughout the recording session to confirm that peak signals were not clipped.

The sampled waveforms were edited into individual sentence files, eliminating silent intervals before and after each waveform. Mean-squared (MS) amplitudes were computed for each sentence waveform. Values ranged from 65.09 to 74.21 dB (*re*: 1 sample unit of a 16-bit digital representation). All waveforms were rescaled to 67 dB to equate initial presentation levels. Only nine sentences were scaled up. The remainder were either scaled down or left unchanged.

After equating the waveform levels, the average long-term spectrum of the sentences was computed. This spectrum served as the target for generation of a spectrally matched masker to be used in measurements of sSRTs in noise. The use of a spectrally matched masker ensures that on average the S/N ratio will be approximately equal at all frequencies. A speech-shaped masker can also increase the sensitivity of the test to changes in speech discrimination. For example, Prosser *et al.* (1991) found steeper intelligibility functions with speech shaped noise versus other types of noise such as traffic noise.

The average long-term spectrum of the sentences was determined by playing back the sentences continuously from the computer through dual cascaded reconstruction filters set to 8 kHz into a dynamic signal analyzer (Hewlett–Packard model 356601). The analyzer was set to measure the long-term average spectrum over a 12-kHz bandwidth, using a Hanning window with continuous rms averaging. Averaging was continued until no further changes in the long-term spectrum were observed. This point was reached after 72 sentences, or approximately 2 min of continuous sentences. A weighted-least-squares filter design program (Nielsen *et al.*, 1990) was used to calculate a 78-coefficient linear phase finite impulse response (FIR) filter based on the spectrum levels at 126 frequencies sampled between 0 and 10 kHz. The filter response matched the target response with an rms error of 1.88 dB (*re*: 1 sample unit) over the 126 frequency samples. Semirandom white noise was synthesized at the original sampling rate, filtered with the FIR filter, and scaled to the same MS amplitude as the speech. Figure 1 shows the average long-term spectrum of the speech, the obtained filter response, and the measured spectrum of the filtered noise. The filter response is shifted upward by 10 dB for comparison purposes.

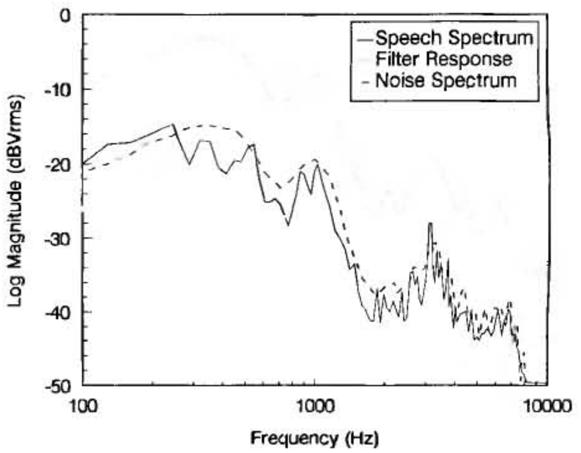

FIG. 1. rms spectrum of sentences from the HINT as calculated from 512 point FFT (solid line). Magnitude response of FIR filter designed to match rms speech spectrum (dotted line). rms spectrum of spectrally matched masking noise (dashed line). The noise spectrum is offset by 10 dB to improve visual clarity.

B. Equating of sentence difficulty

Although the sentences had equal MS amplitude, their intelligibility when presented in the spectrally matched noise would not necessarily be equal. Phonemic content, word familiarity, as well as variations in intonation and level influence intelligibility in noise. Our approach to equating sentence intelligibility was to present the sentences in noise (using the spectrally matched noise) at a fixed S/N ratio to normal-hearing listeners and to measure percent intelligibility, scoring all words in the sentence. The MS level of the sentences was increased if intelligibility was below average, decreased if intelligibility was above average, or left unchanged if intelligibility was approximately at the average. The MS level of the sentence was adjusted approximately 1 dB per 10% difference in intelligibility. This process of scaling to compensate for difficulty was repeated a total of seven times, with the amount of adjustment decreased in each iteration.

1. Method

a. Subjects. A total of 78 native English speaking male and female subjects were paid for their participation. All subjects were screened for normal hearing (15 dB HL from 0.25 to 8 kHz). Their ages ranged from 17 to 45 years with a mean age of 24 years.

b. Apparatus and procedure. Groups of six to eight listeners were tested at a fixed signal-to-noise ratio. All sentence tests were presented diotically under TDH-50 headphones at a fixed S/N ratio in a sound room. Noise was played on a Marantz ND430 stereo cassette recorder and mixed with the speech using a Grason Stadler GSI16 audiometer. The speech was played directly from computer and passed through 8-kHz reconstruction filters with 96-dB/oct attenuation. The noise was played at 72 dB (A), as measured in a 6-cm^3 coupler (B & K Type 4152 Artificial Ear).

The subjects were instructed to listen and repeat aloud whatever was heard or understood. No feedback was pro-

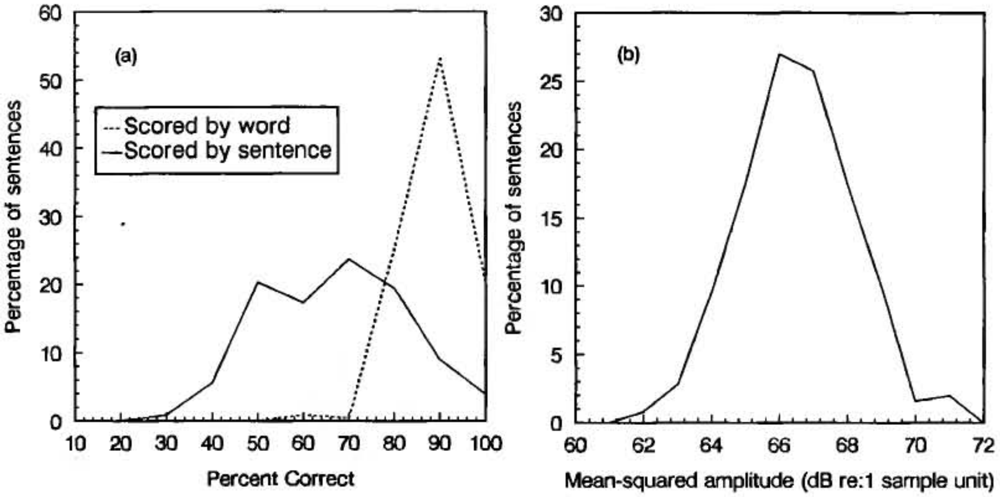

FIG. 2. Panel (a) displays distribution of percent correct scores for 232 sentences approximately equated for difficulty. Sentences were presented in spectrally matched noise at a fixed level. The dashed line is the distribution when scored by words identified correctly, and the solid line is the distribution when scored by sentences correct. Panel (b) displays the distribution of MS amplitudes for 252 sentences after equating their difficulty.

vided, and scoring was done on a word-by-word basis. The first 12 sentences presented to each listener were used as practice. Average percent correct scores for each sentence were calculated from the number of words repeated correctly. During initial tests, only exact repetitions were accepted as correct. The scoring criteria were later relaxed to allow minor variations in articles and verb tenses.

The difference in an individual sentence's percent intelligibility score from the overall mean was used to make an adjustment to the MS level of the sentence (approximately 1 dB per 10% difference). After rescaling of the sentences, the testing was repeated with another set of subjects and the rescaling was repeated. Smaller and smaller amounts of rescaling were applied after each round of testing, with final adjustments after the seventh round of testing of only 0.5 and 0.33 dB. The standard deviation of the distribution of sentence intelligibility scores decreased from 34.2 to 25.1 percent correct as a result of these level adjustments.

The S/N ratio was fixed for all listeners within each testing group, but was varied between groups to verify that extreme sentences were being adequately scaled. A lower fixed S/N ratio was used to reveal differences among the easier sentences (−4 dB S/N), while a higher S/N ratio was used to reveal differences among difficult sentences (−3 dB S/N).

Scoring based upon correct repetition of the entire sentence was also included in the last few analyses, since this scoring method was to be used in the SRT tests. Sentences with acceptable average percent correct word scores that were never 100% correct could be identified in this manner. These problem sentences were subjected to further rescaling or were eliminated if 100% correct was not attained by any listener, even at maximum signal levels. These problem sentences exhibited large level variations, and the lowest level portions were never correctly recognized.

Of the original 336 sentences, 252 were rescaled to approximately equal intelligibility when presented at the same S/N ratio. Figure 2(a) shows the distribution of percent correct scores for 232 of these sentences. The 20 missing sentences were rescaled by very small amounts (i.e., 0.5 to 0.33 dB) to equate their difficulty. These 20 sentences were not submitted to another iteration of intelligibility tests because the small changes were not expected to influence the overall distribution in Fig. 2(a). Note that the mean of this distribution is arbitrarily dependent on the S/N ratio used for equating and that there are no sentences below 20% correct. Figure 2(b) shows the distribution of MS amplitudes (power) of the final 252 sentences. The average MS amplitude is 66.7 dB, with 51% of the sentences falling within ±1 dB of the average.

C. Creation of lists

After matching the sentences for intelligibility, lists of sentences were formed for use in the measurement of sSRTs. The following procedure was used to maximize measurement reliability with lists matched in their phonemic content. The sentences were rewritten in the International Phonetic Alphabet using the Merriam–Webster Pronouncing Dictionary of American English (Kenyon and Knott, 1953) as a guide. The phoneme distribution within the sentence set was determined from these transcriptions. Table I shows the phonemic distribution for the entire set of sentences.

Twenty-one lists of 12 sentences which matched the phonemic distribution of the entire sentence set were formed using a trial-and-error process. For each phoneme in each sentence list, the difference between the target phoneme count (the overall phoneme count divided by the number of lists), as predicted from the overall distribution, and the obtained phoneme count was tabulated. The distribution of the differences between the target and obtained counts for the 43 phonemes in the 21 lists is shown in Fig. 3. A difference of ±1 phoneme was found in 68% of

TABLE I. Phoneme distribution for 252 sentences used in the development of the HINT. Table entries express the rate of occurrence of each phoneme as a percentage of the total phoneme count. The sentences were composed of 4239 phonemes, 2623 consonants (62%) and 1616 vowels (38%).

Consonant distribution					
p	2.2%	θ	0.4%	m	2.7%
b	2.4%	ð	6.5%	n	4.9%
t	5.7%	s	3.6%	ŋ	1.4%
d	3.9%	z	3.7%	l	4.4%
k	3.7%	ʃ	1.3%	w	1.8%
g	1.5%	h	2.5%	j	0.1%
f	2.1%	tʃ	0.8%	r	5.1%
v	0.8%	dʒ	0.3%		
Vowel distribution					
i	2.8%	ɔ	1.5%	ʌ	2.0%
ɪ	5.8%	o	2.0%	aɪ	0.9%
e	2.9%	ʊ	0.6%	aʊ	0.8%
ɛ	2.3%	u	1.0%	ɔɪ	0.4%
æ	1.9%	ɝ	0.5%	ju	0.1%
a	0	ɚ	2.2%		
ɑ	2.2%	ə	8.4%		

the counts. Figure 3 also shows the distribution of difference counts for lists composed of ten sentences, since the final form of the test used lists of ten sentences. Formation of the ten-sentence lists followed the same procedure described above and are listed in the Appendix. Three practice lists were formed from the unused sentences.

III. INTERLIST RELIABILITY

A. Introduction

The next step in the development of the SRT test included experimentation to establish procedural details for SRT measurements and to determine the repeatability, and thus the reliability, of sSRTs measured with different lists. To facilitate administration of the test, an automated, computer-controlled testing procedure was developed. Both the speech and the noise signals were generated from computer hard disk. The signals were then played through programmable reconstruction filters and programmable attenuators. Once levels were calibrated, the computer controlled the speech and noise signals, adjusted the levels according to the adaptive technique, and recorded what materials were presented, as well as the levels of the signals, and the listeners' responses.

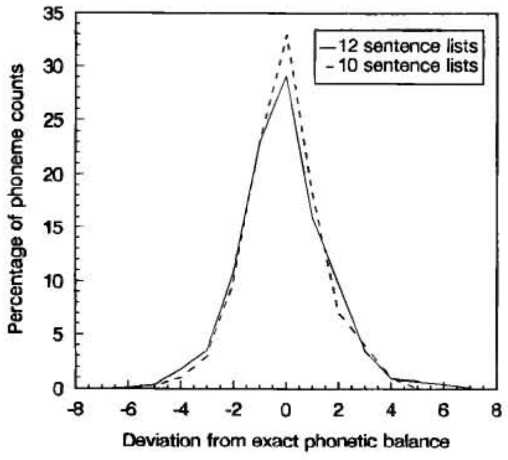

FIG. 3. Deviation of phoneme counts for 12 and 10 sentence lists as a functions of the phoneme count for all 252 sentences. Counts were obtained for 42 English phonemes.

B. Method

1. Subjects

A total of 18 male and female native speakers of English were paid for their participation. One subject was dropped from the analysis because of extraneous noise during testing. All subjects were screened for normal thresholds (15 dB HL from 0.25 to 8 kHz). Their ages ranged from 18 to 43 years, with a mean of 26.8 years.

2. Apparatus

A PC-based, digital processor (Ariel DSP-16) with two dual-cascaded programmable low-pass filters and two programmable attenuators presented the materials. The two channels of output were filtered, attenuated, and then mixed. Levels were set using a Grason–Stadler audiometer (GSI 16). TDH-50 headphones were used as transducers.

3. Design

The sSRTs were measured in four alternating blocks with and without noise, with each block containing five threshold measurements (i.e., five lists). The noise level was fixed at 72 dB(A), *re*: 6 cm^3 coupler. Two practice lists, one in quiet and one in noise, preceded the testing. Sentence order within lists was randomized by the computer, and list order was counterbalanced using a modified Latin-squares design.

4. Procedure

After an initial hearing screening, listeners were asked to listen to each sentence and repeat aloud whatever was heard or understood. An adaptive up–down strategy determined the sentence presentation levels; the first sentence was presented below threshold and was increased in level by 2-dB steps until it was repeated correctly. The subsequent sentences were presented once each, with the presentation level dependent upon the accuracy of the preceding response. Presentation levels were attenuated by 2 dB after a correct response and increased by 2 dB after an incorrect response.

A comparison by the experimenter of the listener's response to a text version of the sentence was used to judge accuracy. The text of the sentences listed specific, small variations in the sentences that were to be allowed as correct responses based upon responses repeatedly made by subjects during the previous testing. These variations were in verb tense ("is" and "was," "are" and "were," and "has" and "had") and in the articles ("a" and "the"). The computer recorded sentence presentation levels, whether

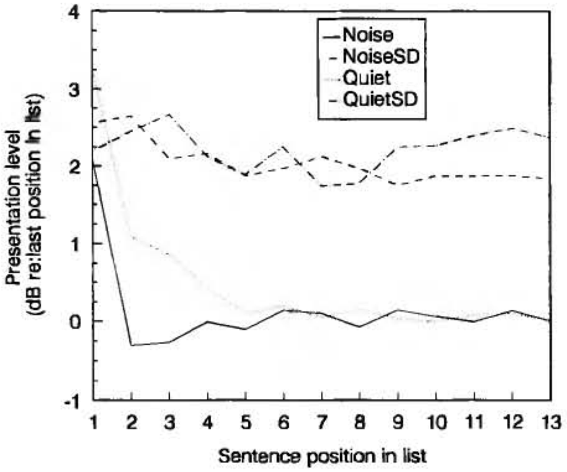

FIG. 4. Means and standard deviations of sentence presentation levels for each list position in a sequence of adaptive threshold measurements. All levels are expressed as dB *re*: the predicted mean level of a 13th sentence in the sequence based upon the response to the 12th sentence. Means and standard deviations are plotted separately for threshold measurements in quiet and in noise.

the listener's response was correct (as determined by the experimenter), as well as the identity of the specific sentence on each trial.

C. Results

The pattern of presentation levels on individual trials was examined first to determine the number of sentences to be included in the SRT calculation. In order to eliminate the effects of subject differences from these analyses, the presentation levels for each subject were expressed in dB *re*: the presentation level on the twelfth (last) trial for each list. These reference calculations for both the mean and standard deviation of presentation levels were performed separately for the noise and quiet presentation conditions. The subject-normed results from these calculations are plotted in Fig. 4. Note that the mean presentation levels stabilize after the fifth sentence, and that the standard deviation of presentation levels changes very little over the entire length of the list. On the basis of these analyses, we concluded that listeners were near their SRT by the fifth trial, and thus that sSRTs could be accurately estimated from the mean presentation levels on the fifth and subsequent trials. All SRT estimates in the remainder of this paper are based on means calculated from the fifth and subsequent sentences in a list, regardless of list length.

The mean S/N ratio at threshold in the noise condition across all subjects was −2.92 dB, with a standard deviation of 0.78 dB. The mean SRT in quiet was 23.91 dB(A), with a standard deviation of 3.45 dB. The greater variability of the quiet thresholds may be due to differences in hearing sensitivity among the subjects in our sample. This issue will be discussed in greater detail below.

We next turned to the question of list equivalence. The mean SRT for each list across all subjects tested with that list was computed and expressed as a deviation score from the mean across all lists and all subjects. Figure 5 shows the mean deviation scores for each list in the quiet and noise test conditions. The error bars show the standard deviation associated with each mean. With only one exception (list 4 in the noise test conditions), all list means fluctuate within 1 dB of the overall mean. An analysis of variance found no significant effect of list type [$F(20.149) = 1.97$, $p > .01$].[1] The grammatical and syntactic level and equivalence of the lists can also be estimated. A relatively simple but repeatable and objective technique for making these estimates exists in the commercial software packages used to grade the reading level of textual materials. One such package (Right Writer 3.1) was used to grade the sentence lists. With the exception of one list rated at second grade reading level, all of the other lists, including the practice lists, were rated at first grade reading level (Stelmachowicz, personal communication). All materials should therefore be easily comprehensible by all adults.

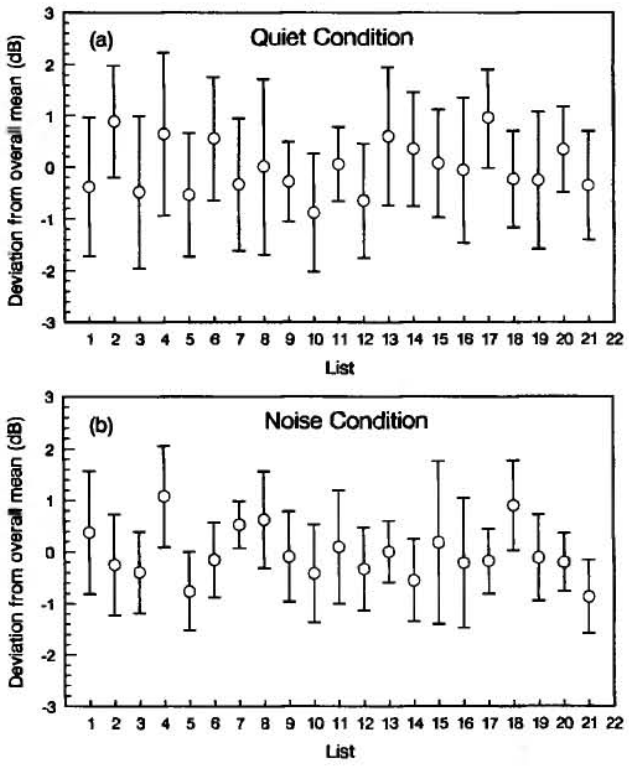

FIG. 5. Differences between means for individual list sSRTs and the mean sSRT across all lists in quiet [panel (a)] and noise [panel (b)]. For each list, $n=8$. Error bars show ±1 standard deviation for each mean.

The reliability and repeatability of the sSRT measures, and consequently their standard error, can be estimated from the standard deviation of differences between repeated sSRT measurements within subjects (Plomp and Mimpen, 1979). These estimates were computed separately for the quiet and noise test conditions. The standard deviation of difference scores in quiet was 1.39 dB, and in noise its value was 1.13 dB. (See Table II.) These values are only slightly higher than Plomp and Mimpen's results with Dutch materials (1.1 dB in quiet and 0.9 dB in noise). Gelfand *et al.* (1988) measured interlist reliability as the signed average of test–retest differences in SRTs obtained

TABLE II. Standard deviations of sSRT difference scores from repeated measurements within subjects, and 95% confidence intervals. Separate entries are given for 10 and 12 sentence lists, and for 1, 2, and 3 lists per measurement.

	12-sentence list			10-sentence list		
	1	2	3	1	2	3
			Quiet			
Standard deviation	1.39	0.92	0.68	1.52	0.99	0.72
Confidence interval	±2.72	±1.80	±1.33	±2.98	±1.94	±1.41
			Noise			
Standard deviation	1.13	0.73	0.57	1.23	0.76	0.61
Confidence interval	±2.21	±1.43	±1.12	±2.41	±1.49	±1.20

with the high-predictability SPIN sentences. These authors reported a mean difference of −0.29 dB, with a standard deviation of 2.27 dB. Thus the present materials are expected to produce somewhat more reliable sSRTs than the adapted SPIN sentences.

Given an estimate of the standard error from the standard deviation of the difference score, the 95% confidence interval or critical difference score for a measurement can also be estimated. Since the SRT is linearly related to the level of the noise when the noise is above the threshold of audibility (e.g., Plomp, 1986), the width of the confidence interval should not be level dependent. The 95% confidence interval or critical difference score is the region bounded by ±1.96 standard deviations about the SRT, with a two-tailed test. The width of the confidence interval depends on the number of lists used to measure the sSRT. Table II reports the width of the 95% confidence intervals estimated for sSRTs measured in quiet and in noise with both 10- and 12-sentence lists. (sSRTs for 10-sentence lists were obtained by eliminating the responses to the 11th and 12th sentences on each list and computing the mean sSRT from the 5th sentence to the end of the list.) Confidence intervals are computed for up to three lists per condition. Note that confidence intervals are only slightly narrower when 12-sentence lists are used, and that increasing the number of lists per condition from one to three reduces the width of the confidence interval by approximately one half, as computed from multiple threshold measurements.

The sensitivity of SRT measurements and thus the ability of the SRT task to reliably detect small threshold shifts can be computed from estimates of the statistical power of the test. Using the computed dispersion measurement (the standard deviation of differences between repeated measures), the probability of detecting a true difference in sSRT can be computed for any sSRT difference (e.g., Hays, 1973, pp. 375–378). Using a normal transformation, statistical power functions describing the relationship between the magnitude of the threshold difference and the probability of correctly rejecting the null hypothesis of no difference were calculated for the noise and quiet conditions, and for one, two, and three lists. Assuming a one-tailed test (i.e., the direction of the threshold difference is stated in the null hypothesis), with a 5% rejection region, the thick curves in Fig. 6 show that the test is more sensitive in the noise conditions, and increases in power are found as scores from multiple lists are averaged to determine the listener's threshold. Trade-offs between test sensitivity and the ability to test a large number of conditions can therefore be made to best suit the testing situation. For example, with three lists per condition, threshold shifts of about 2 dB or more in quiet and 1.5 dB or more in noise

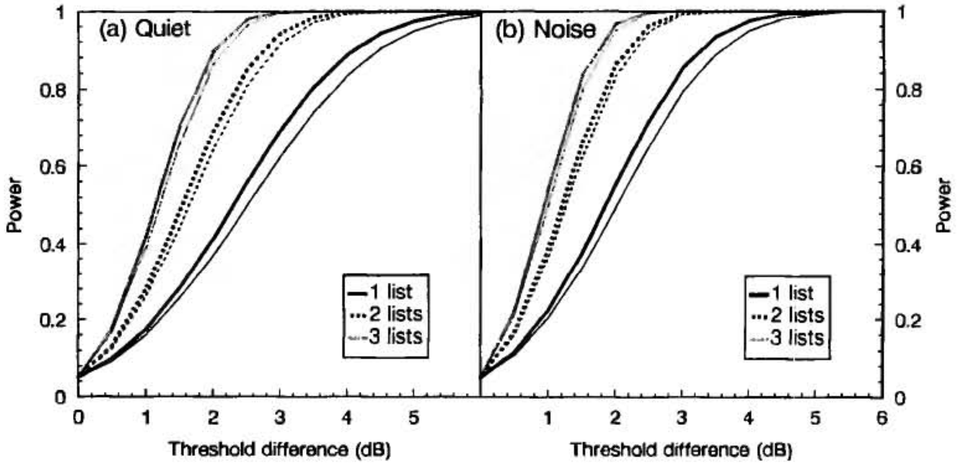

FIG. 6. Panel (a) displays statistical power functions for the HINT in quiet, and panel (b) displays these functions for the HINT in noise. The thick lines show the functions for 12 sentence lists, and the thin lines show the functions for 10 sentence lists. Separate power functions are calculated for measurements taken with 1, 2, and 3 lists.

can be detected with statistical power of 0.9 or greater.

The previous analyses of presentation levels as a function of sentence position in the list, as well as the reliability analyses, suggest that fewer than 12 sentences can be used to obtain reliable sSRTs. We examined this issue by recalculating sSRTs using five different list lengths, from 8 to 12 sentences. We found that sSRTs were essentially unchanged when list length was decreased from 12 to 10 sentences in both quiet and noise, and that standard deviations increased less than 0.1 dB. Therefore, 25 ten-sentence lists were formed to serve as the final version of the HINT. The accuracy of the phonemic balance between lists and the reliability data and power functions were recalculated. (See the dashed line functions in Fig. 3, the narrow line functions in Fig. 6, and Table II.)

D. Discussion

The foregoing results and analyses have shown that reliable SRTs in quiet and noise can be obtained with adaptive testing procedures using short English sentences as stimuli. As few as ten sentences per list will give measurements that are sensitive enough to detect threshold differences of 2.98 dB in quiet, as predicted from the confidence intervals in Table II. Within-subject repeatability of sSRTs is comparable to results obtained by Plomp and Mimpen (1979) with Dutch materials developed specifically for sSRT testing, and is better than the repeatability obtained with American English materials originally developed for other purposes.

Variability of sSRTs in quiet was slightly greater than in noise. We attribute this outcome to the fact that subjects were screened at 15 dB HL for participation in the study, and a number of subjects exhibited pure-tone thresholds near the screening limit. Thus the sSRTs in quiet may reflect differences in hearing sensitivity among subjects, while sSRTs in noise were measured at levels well above the audibility threshold.

The average S/N ratio at threshold in the current study (−2.92 dB) is close to the −2.00-dB value reported by Gelfand *et al.* (1988) for the high-predictability SPIN sentences presented in multitalker babble. Plomp *et al.* (1979), however, obtained sSRTs in spectrally matched noise for sentences developed with a similar procedure at average S/N ratios of −7.3 dB. The Dutch investigators used similar hearing criteria to screen subjects and a dummy head presentation technique (as compared with diotic headphone presentation in the present study). Differences in presentation techniques and instrumentation could influence quiet thresholds; however, no obvious explanation is available for the S/N ratio differences in noise. In the next section we describe a study which examines the origins of the threshold differences in noise for the Dutch and English materials.

E. Comparisons of Dutch and English sentence materials

One aspect of the English sentence materials that could contribute to the elevated S/N ratios at threshold is the range of level variations within the materials. Since the scoring of sentences in the SRT task is based on accurate recognition of the entire sentence, the occurrence of necessary speech information at low levels in the materials could produce elevated S/N ratios at threshold. We have attempted to examine this potential difference between the two sets of materials by obtaining a cassette recording of the Dutch materials (Plomp, personal communication), and by analyzing the distribution of levels in the Dutch and English materials. We have also compared the goodness-of-fit of the spectrally matched noise to the average long-term speech spectrum for both sets of materials.

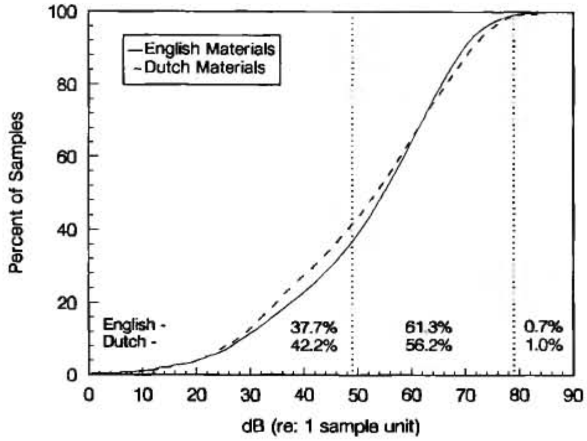

FIG. 7. Cumulative distributions of the instantaneous levels for the HINT sentence materials and for comparable Dutch sentence materials (Plomp *et al.*, 1979). The MS levels of both sets of materials were matched at 67 dB (*re*: 1 sample unit). The vertical dashed lines are positioned 18 dB below and 12 dB above the MS level, bounding the range of speech levels most important for intelligibility as predicted from articulation theory.

1. Method and results

Several minutes of the Dutch sentence materials were sampled from cassette tape into a laboratory computer using the same instrumentation and procedures that were developed for our initial recordings. The long-term rms level of the sampled Dutch materials was computed, and the sampled values were rescaled to match the long-term level of the English sentences. A computer program was developed to compute the cumulative distribution of instantaneous, unsigned magnitudes for both sets of sampled speech waveforms. Thresholding (triggering of the analysis only when the MS power exceeded a set value) was required for the Dutch materials to eliminate low-level hum and tape noise from the analysis.

Figure 7 displays the cumulative level distributions for the Dutch and English materials. The figure shows that a slightly larger proportion of the Dutch materials occur at low levels, i.e., less than 50 dB. These cumulative distributions can be roughly quantified by identifying the range of instantaneous levels above and below the long-term rms level which, according to Articulation Theory (e.g., Humes *et al.*, 1986), contain the most important speech information. Articulation Theory defines the range from 12

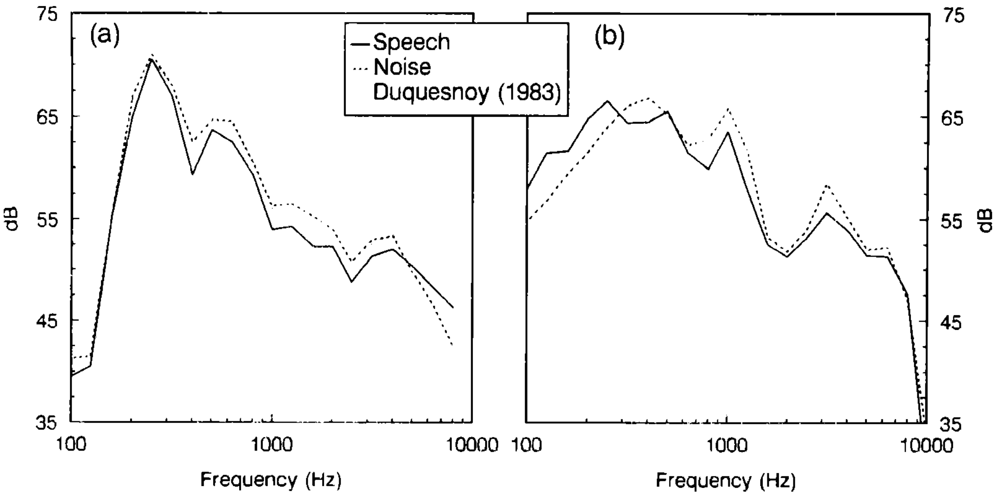

FIG. 8. Panel (a) displays the long-term spectrum of the Dutch sentence materials (solid line) and the spectrally-matched masker (dashed line), as measured in $\frac{1}{3}$ octave bands. Panel (b) shows the same measurements for the HINT materials. The dotted line in panel (a) shows Duquesnoy's (1983) spectral measurements of the Dutch materials for comparison.

dB above the rms level to 18 dB below that level as the most important portion of the speech dynamic range. The vertical dashed lines in Fig. 7 shows the range of levels bounded by rms+12 dB and rms−18 dB. The figure also shows that 61.3% of the English materials and 56.2% of the Dutch materials fall within this range. The percentage of samples falling above and below this range is also shown in the figure. All of these comparisons point to the same conclusion: The Dutch materials are slightly more variable than the English materials. If the lower sSRTs with the Dutch materials were attributable to fewer low level intervals in these materials, then one would expect the Dutch materials to be less variable, not more variable. In short, these analyses tend to rule out the hypothesis that sSRT differences in noise between the English and Dutch materials are caused by differences in level distributions.

Another comparison of the English and Dutch materials is in the fit of the noise spectra to the long-term speech spectra. If the fit differs in the two sets of material, then sSRTs in noise may be influenced. To test this hypothesis, measurements of both the English and Dutch speech and noise were taken with a $\frac{1}{3}$-oct real-time analyzer (General Radio, 1995). The measurements, shown in Fig. 8, are based upon 30 s of continuous speech or noise. Duquesnoy's (1983) original measurements of the Dutch materials are shown with the current measurements in Fig. 8(a). The measurements from the English materials are shown in Fig. 8(b). The noise spectrum of the taped Dutch materials matches Duquesnoy's original measurements up to around 6 kHz, where the spectrum of the taped material drops off.

The spectra of both recordings shows the MS amplitude of the speech materials below the MS amplitude of the noise (1 dB for the Dutch, and 0.3 dB for the English). The test development procedure, common to both sets, matched the sentences for difficulty, not level, thereby producing the differences between speech and noise levels. Level differences are maintained at all but the highest frequencies in the Dutch recordings, and at all but the lowest frequencies in the English materials. These differences are quite small compared to the 4.38 dB sSRT differences between the two sets of materials, and are unlikely to account for the sSRT differences.

Differences in the spectral shape of the Dutch and English materials could contribute to sSRT differences in quiet, since the range of average levels in the Dutch materials is greater than in the English materials. These level differences cannot account for the sSRT differences in noise, however, since the masking noise is well matched to the speech spectrum in both sets of materials. Differences in linguistic entropy between the two sets of materials, i.e., the predictability of their linguistic content (Van Rooij *et al.*, 1991), could also influence the sSRTs. Comparisons of linguistic entropy must await the availability of comparable methods of linguistic analysis for the two languages.

IV. BANDWIDTH STUDY

A. Introduction

The reliability of sSRT measurements will depend to some extent on the audible bandwidth of the speech materials. sSRTs in quiet and in noise are expected to increase as the bandwidth of the speech (and noise) are reduced (e.g., Plomp, 1986). At some point, however, the sSRT will be influenced more by the reduced bandwidth than by the level of the speech. When this point is reached, guessing and response biases will begin to reduce the reliability of SRT measurements. We anticipate that sSRTs will be measured under conditions of reduced audible bandwidth, either with hearing impaired or with limited bandwidth transmission/presentation systems. This study was undertaken to determine the bandwidth over which reliable sSRTs can be obtained.

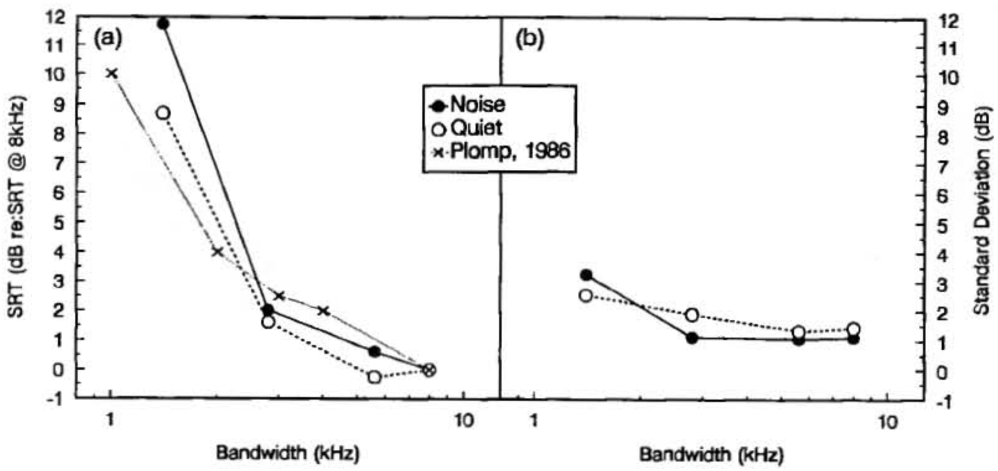

FIG. 9. Panel (a) displays the mean sSRTs in noise (filled circles) and in quiet (open circles) as a function of bandwidth. Means are expressed in dB *re*: mean sSRT in the full bandwidth (8 kHz) condition. Plomp's (1986) mean sSRTs in noise are displayed for comparison. Panel (b) displays the average standard deviation of differences in repeated measures within subject as a function of bandwidth.

B. Method

1. Subjects

A total of 24 native English speaking male and female subjects were paid to participate in the experiment. All subjects were screened for normal thresholds (15 dB HL from 0.25 to 8 kHz). Their ages ranged from 21 to 45 years, with a mean of 27.9 years.

2. Apparatus and design

Subjects were tested with the same equipment and test materials as in the reliability study. Subjects were presented two sets of five lists [one set in quiet and one set in 72 dB(A) noise] at the full bandwidth (8 kHz) and two sets of five lists at one of three reduced bandwidths. We selected bandwidths based on the upper band edges of the 1-, 2-, and 4-kHz octave bands: 1.4, 2.8, and 5.6 kHz, respectively. These selections allow us to describe the results in terms of the octave band contributions to sSRT and reliability. The reconstruction filters (96-dB/oct roll-off) were used to create the reduced bandwidth materials. The order and choice of test conditions was counterbalanced across subjects, and thresholds were measured using the same adaptive technique described above. Listeners were tested in the full bandwidth condition and in one of the limited bandwidth conditions. Thus the comparison of full versus reduced bandwidth was made within subjects, and comparisons of the various reduced bandwidth conditions was made between subjects. Separate analyses of variance in the quiet and noise conditions were used to make these comparisons.

C. Results

Figure 9(a) displays the mean sSRTs in quiet and noise for each bandwidth, along with sSRTs in noise from Plomp (1986) for comparison with previous research. All sSRTs are expressed in dB *re*: threshold in the full bandwidth condition to facilitate comparisons. sSRTs in the reduced bandwidth conditions were significantly higher than in the full bandwidth condition, $F(1,20)=53.28$, $p<0.01$. There were also significant differences among the sSRTs in the reduced bandwidth conditions, $F(2,20)=21.46$, $p<0.01$, with sSRTs ranging from 0 to 8.5 dB higher than in the full bandwidth condition.

A similar pattern of results is seen for the sSRTs in noise. sSRTs in the reduced bandwidth conditions were again significantly higher than in the full bandwidth condition, $F(2,21)=275.80$, $p<0.01$. sSRTs in the reduced bandwidth conditions differed significantly, $F(2,21)=92.96$, $p<0.01$), ranging from approximately 0.5 to 12 dB higher than the full bandwidth condition.

Each reduced bandwidth condition from the quiet and noise experiments was compared with the full bandwidth condition in separate analyses of variance. These analyses revealed that sSRTs did not increase significantly when the speech and noise were low-pass filtered at the upper edge of the 4-kHz octave band (i.e., 5.6 kHz), eliminating the 8-kHz octave band. Low-pass filtering at the upper edge of the 2-kHz octave band, eliminating the 4-kHz octave band, however, did significantly increase sSRTs. These results are similar to Plomp's (1986) findings above about 2.5 kHz. To summarize, both the present study and Plomp's data indicate that sSRTs in quiet and noise increase by about 3 dB as bandwidth is reduced to around 2–2.5 kHz. As bandwidth decreases from 2 to 1 kHz, sSRTs increase rapidly by another 7–9 dB.

The reliability of sSRTs as a function of bandwidth can be estimated from the standard deviation of the mean difference between repeated measures in each condition. These standard deviations are plotted as a function of bandwidth for the quiet and noise conditions in Fig. 9(b). In the full bandwidth conditions, standard deviations ranged between 1–1.5 dB, and did not change when the 8-kHz octave band was eliminated. When the 4-kHz octave band was eliminated, standard deviations increased by about 0.5 dB in quiet and remained unchanged in noise. Only when the 2-kHz octave band was also eliminated, were there substantial increases in the standard deviations to about 2 dB in quiet and 3 dB in noise.

D. Discussion

The results of this experiment show the effects of signal and noise bandwidth on the mean and standard deviation of sSRTs measured in quiet and in noise. The observed pattern of increasing sSRTs with reductions in bandwidth replicates approximately the previous research by Plomp (1986) with Dutch sentence materials. Elimination of the 8-kHz octave band has little influence on the sSRT, while elimination of the 4- and 8-kHz octave bands elevated thresholds by about 2 dB. Elimination of the 2-, 4-, and 8-kHz bands, however, elevates thresholds about 10 dB. Perhaps more importantly, the reliability of the sSRTs—as estimated from the standard deviation of differences in repeated measures—is substantially degraded only after the bandwidth drops below about 2 kHz. In other words, the reliability, and thus the sensitivity of the sSRT test is fairly robust amidst variations in audible bandwidth associated with hearing impairment and prosthetic devices used to remediate this impairment. This is not to say that increased variability is not expected with the hearing impaired, but the test has the potential for reliability close to that found with normal-hearing subjects.

V. SUMMARY

A recorded test for measurements of sentence intelligibility in quiet and in noise has been developed. We have titled this test the hearing in noise test (HINT). The test is composed of recordings of 250 sentences produced by a male speaker. The sentences were derived from the Bamford–Kowal–Bench British sentence materials and were rewritten in American English. The sentences were also equated for naturalness, length, and intelligibility. The sentences are cast into 25 phonemically matched and balanced lists with ten sentences per list.

The sentence lists in the HINT are intended for adaptive measurement of speech reception thresholds in quiet or in spectrally matched noise. The adaptive procedure avoids nonlinear ceiling and floor effects customarily associated with intelligibility test given at fixed presentation levels, and optimizes the statistical efficiency of the test. A threshold measurement with a single list usually takes less than two minutes, contributing to the practical 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. The 95% confidence interval for ten-sentence lists is estimated at ±2.41 dB for thresholds measured in noise, and ±2.98 dB for thresholds in quiet. Estimated confidence intervals in quiet are slightly larger than in noise because of the variability in hearing levels among subjects, even though they were screened at 15 dB HL from 0.25 to 8 kHz. Improved confidence intervals can be obtained with multiple lists. The reliability of the HINT compares favorably with the reliability of the Dutch materials (Plomp *et al.*, 1979), and is better than the reliability of the SPIN sentences (Gelfand *et al.*, 1988). Sensitivity of the HINT is given by statistical power curves calculated for presentations of single and multiple lists.

Average thresholds with the HINT were 23.91 dB(A) in quiet, and 69.08 dB(A) in 72 dB(A) noise (i.e., −2.92 dB S/N). These threshold values are not intended to define norms for the HINT because of limitations in sample size. Subsequent research will establish such norms. Comparisons of these thresholds with Plomp and Mimpen's (1979) results with comparable Dutch sentences showed that the Dutch thresholds were several dB lower. Analyses of the Dutch and English sentences and their respective noise maskers did not reveal the source of these threshold differences.

The effect of bandwidth on sSRTs and their reliability were also examined. sSRTs in quiet and noise began to increase when the 4- and 8-kHz octave bands were eliminated by low-pass filtering. However, the largest increases in sSRT occurred when the 2-kHz octave band was eliminated, and the bandwidth dropped below 2.5 kHz. These results are consistent with previous research. Reliability of the HINT also is not degraded substantially until bandwidth drops below 2.5 kHz.

The HINT provides an accurate, reliable, and efficient method of measuring speech intelligibility in noise. Norms for the HINT based on a large sample of normal-hearing subjects will be established in future research (Nilsson *et al.*, 1992). These norms can be used to assess directly the impact of hearing impairment on communication handicap in quiet and in noise.

ACKNOWLEDGMENTS

The authors would like to thank Lena Wong for assistance in the data collection. We would also like to thank Ed Wu for the software design, Sriram Jayaraman, and John Wygonski for their digital signal processing expertise and assistance, and Jeffrey Danhauer for his useful comments on an earlier version of this manuscript.

APPENDIX

Below are lists of HINT sentences including the variations in response that are allowed (the word used in the recording is underlined). It should be emphasized that the results presented in the paper hold only for a recording of these lists by a male speaker. Copies of the materials are available for purchase from the House Ear Institute.

List 1

1. (A/the) boy fell from (a/the) window.
2. (A/the) wife helped her husband.
3. Big dogs can be dangerous.
4. Her shoes (are/were) very dirty.
5. (A/the) player lost (a/the) shoe.
6. Somebody stole the money.
7. (A/the) fire (is/was) very hot.
8. She's drinking from her own cup.

9. (A/the) picture came from (a/the) book.
10. (A/the) car (is/was) going to fast.

List 2

1. (A/the) boy ran down (a/the) path.
2. Flowers grow in (a/the) garden.
3. Strawberry jam (is/was) sweet.
4. (A/the) shop closes for lunch.
5. The police helped (a/the) driver.
6. She looked in her mirror.
7. (A/the) match fell on (a/the) floor.
8. (A/the) fruit came in (a/the) box.
9. He really scared his sister.
10. (A/the) tub faucet (is/was) leaking.

List 3

1. They heard (a/the) funny noise.
2. He found his brother hiding.
3. (A/the) dog played with (a/the) stick.
4. (A/the) book tells (a/the) story.
5. The matches (are/were) on (a/the) shelf.
6. The milk (is/was) by (a/the) front door.
7. (A/the) broom (is/was) in (a/the) corner.
8. (A/the) new road (is/was) on (a/the) map.
9. She lost her credit card.
10. (A/the) team (is/was) playing well.

List 4

1. (A/the) little boy left home.
2. They're going out tonight.
3. (A/the) cat jumped over (a/the) fence.
4. He wore his yellow shirt.
5. (A/the) lady sits in her chair.
6. He needs his vacation.
7. She's washing her new silk dress.
8. (A/the) cat drank from (a/the) saucer.
9. Mother opened (a/the) drawer.
10. (A/the) lady packed her bag.

List 5

1. (A/the) boy did (a/the) handstand
2. They took some food outside.
3. The young people (are/were) dancing.
4. They waited for an hour.
5. The shirts (are/were) in (a/the) closet.
6. They watched (a/the) scary moving.
7. The milk (is/was) in (a/the) pitcher.
8. (A/the) truck drove up (a/the) road.
9. (A/the) tall man tied his shoes.
10. (A/the) letter fell on (a/the) floor.

List 6

1. (A/the) silly boy (is/was) hiding.
2. (A/the) dog growled at the neighbors.
3. (A/the) tree fell on (a/the) house.
4. Her husband brought some flowers.
5. The children washed the plates.
6. They went on vacation.
7. Mother tied (a/the) string too tight.
8. (A/the) mailman shut (a/the) gate.
9. (A/the) grocer sells butter.
10. (A/the) baby broke his cup.

List 7

1. The cows (are/were) in (a/the) pasture.
2. (A/the) dishcloth (is/was) soaking wet.
3. They (have/had) some chocolate pudding.
4. She spoke to her eldest son.
5. (An/the) oven door (is/was) open.
6. She's paying for her bread.
7. My mother stirred her tea.
8. He broke his leg again.
9. (A/the) lady wore (a/the) coat.
10. The cups (are/were) on (a/the) table.

List 8

1. (A/the) ball bounced very high.
2. Mother cut (a/the) birthday cake.
3. (A/the) football game (is/was) over.
4. She stood near (a/the) window.
5. (A/the) kitchen clock (is/was) wrong.
6. The children helped their teacher.
7. They carried some shopping bags.
8. Someone (is/was) crossing (a/the) road.
9. She uses her spoon to eat.
10. (A/the) cat lay on (a/the) bed.

List 9

1. School got out early today.
2. (A/the) football hit (a/the) goalpost.
3. (A/the) boy ran away from school.
4. Sugar (is/was) very sweet.
5. The two children (are/were) laughing.
6. (A/the) fire truck (is/was) coming.
7. Mother got (a/the) sauce pan.
8. (A/the) baby wants his bottle.
9. (A/the) ball broke (a/the) window.
10. There (is/was) a bad train wreck.

List 10

1. (A/the) boy broke (a/the) wooden fence.
2. (An/the) angry man shouted.
3. Yesterday he lost his hat.
4. (A/the) nervous driver got lost.
5. (A/the) cook (is/was) baking (a/the) cake.
6. (A/the) chicken laid some eggs.
7. (A/the) fish swam in (a/the) pond.
8. They met some friends at dinner.
9. (A/the) man called the police.
10. (A/the) truck made it up (a/the) hill.

List 11

1. (A/the) neighbor's boy (has/had) black hair.

2. The rain came pouring down.
3. (An/the) orange (is/was) very sweet.
4. He took the dogs for a walk.
5. Children like strawberries.
6. Her sister stayed for lunch.
7. (A/the) train (is/was) moving fast.
8. Mother shut (a/the) window.
9. (A/the) bakery (is/was) open.
10. Snow falls in the winter.

List 12

1. (A/the) boy went to bed early.
2. (A/the) woman cleaned her house.
3. (A/the) sharp knife (is/was) dangerous.
4. (A/the) child ripped open (a/the) bag.
5. They had some cold cuts for lunch.
6. She's helping her friend move.
7. They ate (a/the) lemon pie.
8. They (are/were) crossing (a/the) street.
9. The sun melted the snow.
10. (A/the) little girl (is/was) happy.

List 13

1. She found her purse in (a/the) trash.
2. (A/the) table (has/had) three legs.
3. The children waved at (a/the) train.
4. Her coat (is/was) on (a/the) chair.
5. (A/the) girl (is/was) fixing her dress.
6. It's time to go to bed.
7. Mother read the instructions.
8. (A/the) dog (is/was) eating some meat.
9. Father forgot the bread.
10. (A/the) road goes up (a/the) hill.

List 14

1. The fruit (is/was) on the ground.
2. They followed (a/the) garden path.
3. They like orange marmalade.
4. There (are/were) branches everywhere.
5. (A/the) kitchen sink (is/was) empty.
6. The old gloves (are/were) dirty.
7. The scissors (are/were) very sharp.
8. (A/the) man cleaned his suede shoes.
9. (A/the) raincoat (is/was) dripping wet.
10. It's getting cold in here.

List 15

1. (A/the) house (has/had) nine bedrooms.
2. They're shopping for school clothes.
3. They're playing in (a/the) park.
4. Rain (is/was) good for the trees.
5. They sat on (a/the) wooden bench.
6. (A/the) child drank some fresh milk.
7. (A/the) baby slept all night.
8. (A/the) salt shaker (is/was) empty.
9. (A/the) policeman knows the way.
10. The buckets fill up quickly.

List 16

1. He played with his toy train.
2. They're watching (a/the) cuckoo clock.
3. Potatoes grow in the ground
4. (A/the) girl ran along (a/the) fence.
5. (A/the) dog jumped on (a/the) chair.
6. They finished dinner on time.
7. He got mud on his shoes.
8. They're clearing (a/the) table.
9. Some animals sleep on straw.
10. The police cleared (a/the) road.

List 17

1. Mother picked some flowers.
2. (A/the) puppy played with (a/the) ball.
3. (An/the) engine (is/was) running.
4. (An/the) old woman (is/was) at home.
5. They're watching (a/the) train go by.
6. (An/the) oven (is/was) too hot.
7. They rode their bicycles.
8. (A/the) big fish got away.
9. They laughed at his story.
10. They walked across the grass.

List 18

1. (A/the) boy (is/was) running away.
2. (A/the) towel (is/was) near (a/the) sink.
3. Flowers can grow in (a/the) pot.
4. He's skating with his friend.
5. (A/the) janitor swept (a/the) floor.
6. (A/the) lady washed (a/the) shirt.
7. She took off her fur coat.
8. The match boxes (are/were) empty.
9. (A/the) man (is/was) painting (a/the) sign.
10. (A/the) dog came home at last.

List 19

1. (A/the) painter uses (a/the) brush.
2. (A/the) family bought (a/the) house.
3. Swimmers can hold their breath.
4. She cut (a/the) steak with her knife.
5. They're pushing an old car.
6. The food (is/was) expensive.
7. The children (are/were) walking home.
8. They (have/had) two empty bottles.
9. Milk comes in (a/the) carton.
10. (A/the) dog sleeps in (a/the) basket.

List 20

1. (A/the) clown (has/had) (a/the) funny face.
2. The bath water (is/was) warm.
3. She injured four of her fingers.
4. He paid his bill in full.
5. They stared at (a/the) picture.
6. (A/the) driver started (a/the) car.
7. (A/the) truck carries fresh fruit.
8. (A/the) bottle (is/was) on (a/the) shelf.

9. The small tomatoes (are/were) green.
10. (A/the) dinner plate (is/was) hot.

List 21

1. They're running past (a/the) house.
2. He's washing his face with soap.
3. (A/the) dog's chasing (a/the) cat.
4. (A/the) milkman drives (a/the) small truck.
5. (A/the) bus leaves before (a/the) train.
6. (A/the) baby (has/had) blue eyes.
7. (A/the) bag fell off (a/the) shelf.
8. They (are/were) coming for dinner.
9. They wanted some potatoes.
10. They knocked on (a/the) window.

List 22

1. (A/the) girl came into (a/the) room.
2. (A/the) field mouse found (a/the) cheese.
3. They're buying some fresh bread.
4. (A/the) machine (is/was) noisy.
5. (A/the) rice pudding (is/was) ready.
6. They had a wonderful day.
7. (An/the) exit (is/was) well lit.
8. (A/the) train stops at (a/the) station.
9. He (is/was) sucking his thumb.
10. (A/the) big boy kicked the ball.

List 23

1. The paint dripped on the ground.
2. (A/the) towel fell on (a/the) floor.
3. (A/the) family likes fish.
4. The bananas (are/were) too ripe.
5. He grew lots of vegetables.
6. She argues with her sister.
7. (A/the) kitchen window (is/was) clean.
8. He hung up his raincoat.
9. (A/the) mailman brought (a/the) letter.
10. (A/the) mother heard (a/the) baby.

List 24

1. (A/the) waiter brought (a/the) cream.
2. (A/the) teapot (is/was) very hot.
3. (An/the) apple pie (is/was) good.
4. (A/the) jelly jar (is/was) full.
5. (A/the) girl (is/was) washing her hair.
6. (A/the) girl played with (a/the) baby.
7. (Athe) cow (is/was) milked every day.
8. They called an ambulance.
9. They (are/were) drinking coffee
10. He climbed up (a/the) ladder.

List 25

1. (A/the) boy slipped on the stairs.
2. New neighbors (are/were) moving in.
3. (A/the) girl caught (a/the) head cold.
4. His father will come home soon.
5. (A/the) bus stopped suddenly.
6. He (is/was) washing his car.
7. (A/the) cat caught (a/the) little mouse.
8. They broke all the brown eggs.
9. (A/the) candy shop (is/was) empty.
10. (A/the) lady went to (a/the) store.

Practice list 1

1. (A/the) boy got into trouble.
2. The yellow pears taste good.
3. (A/the) front yard (is/was) pretty.
4. (An/the) old man (is/was) worried.
5. The pond water (is/was) dirty.
6. (A/the) rancher (has/had) (a/the) bull.
7. The ground (is/was) very hard.
8. They painted (a/the) wall white.
9. Dad stopped to pick some pears.
10. She made her bed and left.
11. He cut his index finger.
12. (A/the) baby (is/was) on (a/the) rug.

Practice list 2

1. Men normally wear long pants.
2. (A/the) house (has/had) (a/the) nice garden.
3. (A/the) little girl (is/was) shouting.
4. (A/the) driver waited for me.
5. The three girls (are/were) listening.
6. (An/the) ice cream (is/was) melting.
7. She bumped here head on (a/the) door.
8. (An/the) apple pie (is/was) baking.
9. She's calling her daughter.
10. (A/the) park (is/was) near (a/the) road.
11. (A/the) baby (is/was) pretty.
12. They washed in cold water.

Practice list 3

1. (A/the) boy forgot his book.
2. (A/the) mouse ran into (a/the) hole.
4. He closed his eyes and jumped.
5. (A/the) floor looks clean and shiny.
6. She writes to her friend daily.
7. The two farmers (are/were) talking.
8. Father paid at (a/the) gate.
9. They're climbing (an/the) old oak tree.
10. The sky (is/was) very blue.
11. (A/the) black dog (is/was) hungry.
12. They lost all their money.

[1] This ANOVA was for the quiet condition. A significant effect of list was found in the noise condition [$f(20,149) = 2.25$, $p < 0.01$], but post-hoc analysis revealed the difference was only significant between the two extreme lists (lists 4 and 21 with a mean difference in signal-to-noise ratio of 2.2 dB). With 210 comparisons possible when calculating all pairwise comparison of 21 lists, 2 significant differences are expected by chance at the 0.01 level. This single difference, along with the small number of scores for each list (from 7 to 10 scores per list), and the lack of an effect in quiet led us to conclude that there are no reliable list differences.

American Academy of Otolaryngology Committee on Hearing and Equilibrium, and the American Council of Otolaryngology Committee on the Medical Aspects of Noise. (**1979**). "Guide for the evaluation of hearing handicap," J. Am. Med. Assoc. **241**(19), 2055–2059.

Bench, J., and Bamford, J. (Eds.) (**1979**). *Speech-hearing Tests and the Spoken Language of Hearing-impaired Children* (Academic, London).

Bilger, R. C., Nuentzeq, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (**1984**). "Standardization of a test of Speech Perception in Noise," J. Speech Hear. Res. **27**, 32–48.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., and Kishon-Rabin, L. (**1988**). "Voice fundamental frequency as an auditory supplement to the speechreading of sentences," Ear Hear. **9**(6), 306–312.

Carhart, R. (**1946**). "Monitored live voice as a test of auditory acuity," J. Acoust. Soc. Am. **17**, 339–349.

Cox, R. M., Alexander, G. C., and Gilmore, C. (**1987**). "Development of the Connected Speech Test (CST)," Ear Hear. **8**(5), 119s–126s.

Dirks, D. D., Morgan, D. E., and Dubno, J. R. (**1982**). "A procedure for quantifying the effects of noise on speech recognition," J. Speech Hear. Disord. **47**, 114–123.

Dubno, J. R., Dirks, D. D., and Morgan, D. E. (**1984**). "Effects of age and mild hearing loss on speech recognition in noise," J. Acoust. Soc. Am. **76**, 87–96.

Duquesnoy, A. J. (**1983**). "The intelligibility of sentences in quiet and in noise in aged listeners," J. Acoust. Soc. Am. **74**, 1136–1144.

Gelfand, S. A., Ross, L., and Miller, S. (**1988**). "Sentence reception in noise from one versus two sources: Effects of aging and hearing loss," J. Acoust. Soc. Am. **83**, 248–256.

Hagerman, D. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.

Hagerman, D. (**1984**). "Clinical measurements of speech reception thresholds in noise," Scand. Audiol. **13**, 57–63.

Hays, W. L. (**1973**). *Statistics for the Social Sciences* (Holt, Reinhart, and Winston, Inc., New York), second edition.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (**1952**). "Development of materials for speech audiometry," J. Speech Hear. Disord. **17**, 321–337.

Humes, L. E., Dirks, D. D., Bell, T. S., Ahlstrom, C., and Kincaid, G. E. (**1986**). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **29**, 447–462.

Kalikow, D. N., Stevens, K. N., and Elliot, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.

Kenyon, J. S., and Knott, T. A. (**1953**). *A Pronouncing Dictionary of American English* (Merriam, Springfield, MA).

Laurence, R. B., Moore, B. C. J., and Glasberg, B. R. (**1983**). "A comparison of behind-the-ear high-fidelity linear hearing aids and two-channel compression aids, in the laboratory and in everyday life," Br. J. Audiol. **17**, 31–48.

Levitt, H. (**1978**). "Adaptive testing in Audiology," Scand. Audiol. Suppl. **6**, 241–291.

Macleod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**, 131–141.

Macleod, A., and Summerfield, Q. (**1990**). "A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use," Br. J. Audiol. **24**, 29–43.

Nielsen, L. B., Wu., E., Hoffman, M. W., Buckley, K., and Soli, S. D. (**1990**). "Design and evaluation of FIR filters for digital hearing aids with arbitrary amplitude and phase response," J. Acoust. Soc. Am. Suppl. **87**, S24.

Nilsson, M. J., Sullivan, J., and Soli, S. D. (**1991a**). "Validation of a speech intelligibility test using SRTs for hearing aid research," J. Acoust. Soc. Am. **89**, 1960(A).

Nilsson, M. J., Sullivan, J., and Soli, S. D. (**1991b**). "Measurement and prediction of hearing handicap using an additive noise model," J. Acoust. Soc. Am. **90**, 326(A).

Nilsson, M. J., Gelnett, D., Sullivan, J., Soli, S. D., and Goldberg, R. L. (**1992**). "Norms for the Hearing In Noise Test: The influence of spatial separation, hearing loss, and English language experience on speech reception thresholds," J. Acoust. Soc. Am. **92**, 385(A).

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

Prosser, S., Turrini, M., and Arslan, E. (**1991**). "Effects of different noises on speech discrimination by the elderly," Acta Oto-Laryngol. (Stockholm), **s476**, 136–142.

Smoorenburg, G. F. (**1992**). "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram," J. Acoust. Soc. Am. **91**, 421–437.

Van Rooij, J. C. G. M., and Plomp, R. (**1991**). "The effects of linguistic entropy on speech perception in noise in young and elderly listeners," J. Acoust. Soc. Am. **90**, 2985–2991.