Using listening difficulty ratings of conditions for speech communication in rooms

Hiroshi Sato[a)]
Institute for Research in Construction, National Research Council, Ottawa, K1A 0R6 Canada

John S. Bradley
Institute for Research in Construction, National Research Council, Ottawa, K1A 0R6 Canada

Masayuki Morimoto
Environmental Acoustics Laboratory, Faculty of Engineering, Kobe University, Rokko, Nada, Kobe 657-8501 Japan

(Received 23 March 2004; revised 2 November 2004; accepted 24 November 2004)

The use of listening difficulty ratings of speech communication in rooms is explored because, in common situations, word recognition scores do not discriminate well among conditions that are near to acceptable. In particular, the benefits of early reflections of speech sounds on listening difficulty were investigated and compared to the known benefits to word intelligibility scores. Listening tests were used to assess word intelligibility and perceived listening difficulty of speech in simulated sound fields. The experiments were conducted in three types of sound fields with constant levels of ambient noise: only direct sound, direct sound with early reflections, and direct sound with early reflections and reverberation. The results demonstrate that (1) listening difficulty can better discriminate among these conditions than can word recognition scores; (2) added early reflections increase the effective signal-to-noise ratio equivalent to the added energy in the conditions without reverberation; (3) the benefit of early reflections on difficulty scores is greater than expected from the simple increase in early arriving speech energy with reverberation; (4) word intelligibility tests are most appropriate for conditions with signal-to-noise (S/N) ratios less than 0 dBA, and where S/N is between 0 and 15-dBA S/N, listening difficulty is a more appropriate evaluation tool. [DOI: 10.1121/1.1849936]

PACS numbers: 43.55.Hy, 43.71.Gv [NX] Pages: 1157–1167

I. INTRODUCTION

There are many situations in rooms where the results of speech intelligibility tests would suggest that conditions are reasonably acceptable, with intelligibility scores of 90% or greater and signal-to-noise ratios (S/N) above 0 dBA. Such conditions are actually very common. For example, Pearsons[1] found speech-to-noise ratios in public spaces varying from 0 dBA in an aircraft cabin to 15 dBA in a classroom. It is easy to appreciate that conditions of 0-dBA S/N are quite different than those with a 10- or 15-dBA S/N, even though word intelligibility scores for both conditions are similar and quite close to 100%. Such results disguise the fact that in these conditions, speech intelligibility is only possible with a great amount of extra effort by the listener. In fact, it is quite difficult to understand speech in many of these conditions and it seems incorrect to suggest that they represent good conditions for speech communication.

The authors have previously developed the use of subjective ratings of listening difficulty as a better indicator of the quality of acoustical conditions in rooms for speech communication[2] using speech tests in Japanese. Listening difficulty ratings were better able to discriminate among conditions above 0-dB S/N than could speech intelligibility scores. Several other approaches have been used previously to assess speech communication performance instead of speech recognition scores. For example, subjective ratings of the ''easiness'' of speech recognition or ''ease of listening'' were considered.[3–5] Because ''ease'' ratings were measured using paired comparison tests and category scaling methods, they result in a relative scale, and one cannot say how easy is good enough, or when a low rating corresponds to unacceptably bad acoustical conditions. Apoux[6] rated ''ease of listening'' with reaction time for consonants and other studies have measured ''listening effort'' such as those by Downs,[7] and Hicks and Tharpe[8] using a dual-task paradigm. In these studies, the probe reaction time was measured after the word recognition task. The reaction time was treated as listening effort. This is suitable as a clinical technique to assess the effectiveness of hearing-aid devices and/or the application of signal-processing techniques to such devices. However, it is difficult to measure reaction time for groups of typical listeners in real rooms, the focus of this study. This type of measure also has the problem that there is no reaction time value that can be said to be good enough or that corresponds to no listening difficulty.

In rooms, early-arriving reflections have been shown to be particularly important for good speech communication. Lochner and Burger[9] and others have provided a solid basis for the importance of such early reflections. However, con-

[a)]Presently at Institute for Human Science & Biomedical Engineering, National Institute of Advanced Industrial Science and Technology, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan. Electronic mail: sato.hiro@aist.go.jp

ventional room acoustics design is still usually based only on consideration of appropriate reverberation times.[10] A recent study by the authors demonstrated the expected improvements to speech intelligibility scores due to early reflections and indicated that in realistic situations early reflection energy in real rooms is equivalent to increasing the level of the direct sound by up to 9 dB.[11] That is, the effective useful speech level could be as much as 9 dB greater than the direct sound due to the beneficial effects of early reflections of speech sounds. Of course, later-arriving speech sounds are usually found to be detrimental to the intelligibility of speech.[9]

The present study examines the use of listening difficulty ratings to rate speech transmission performance as details of the sound-field components were varied. This included examining whether the benefits of early-arriving reflections to listening difficulty ratings are similar to their effect on speech intelligibility scores. This was done using two different types of tests to assess word intelligibility and perceived difficulty of listening to speech in simulated sound fields that were similar to those used in a previous study.[11] Although sound fields can be described by the complete details of impulse responses, the perceptually important aspects can be more simply understood by considering the speech levels associated with three basic components: the direct sound, the early reflections, and the later-arriving reverberant speech sound. In this new work, sound fields were varied by varying each of these components as a group and without changing the detailed make-up of each group such as the details of individual early reflections. It was intended that the effects of changes to each of these three component groups would be indicative of the effects of similar changes in a wide range of rooms.

The first experiment used sound fields that included either, only direct sound (D), or direct sound with early reflections ($D+E$), and combined with two different levels of steady ambient noise. The second experiment used three types of sound fields: direct sound only (D), direct sound with reverberation ($D+\text{Rev.}$), and direct sound with early reflections and reverberation ($D+E+\text{Rev.}$), all with a constant level of ambient noise. Additionally, paired comparison tests were used to confirm the significance of the differences among some of the sound fields in the second experiment.

The main goal of the new work reported in this paper was to confirm and extend previous results[2] that indicated listening difficulty is a better rating of the quality of conditions for speech communication than word recognition test scores in the range of S/N that occurs most frequently in actual rooms. This new work extends the previous work by using native and non-native speakers with a range of language skills and used speech tests in English rather than Japanese. It also explored the effects of systematic variations in the sound fields representative of conditions commonly found in real rooms including varied levels of early-arriving reflections. The new work focuses on conditions with S/N of 0 dBA or greater, whereas our previous work[11] considered the benefits of early reflections on intelligibility scores in more adverse conditions with S/N of 5 dBA and less.

II. LISTENING TESTS WITH AMBIENT NOISE AND VARIED EARLY REFLECTIONS

A. Method

1. Sound-field simulation procedures

All simulated sound fields were produced using a seven-channel electro-acoustic system with loudspeakers arranged around the listener in an anechoic room at the Institute for Research in Construction, National Research Council Canada. The seven loudspeakers were located at a distance of 1.7 m from the listener and at varied angular locations relative to the listener to simulate early reflections from various angles. Each of the seven channels of electronics included programmable digital equalizers with time delays that could all be changed under computer control via a MIDI interface. The loudspeaker responses were corrected to be flat ± 3 dB from 80 Hz to 12 kHz.

The loudspeaker located directly in front of the listener produced the simulated direct sound (first arriving sound). The other six loudspeakers each produced one early reflection. The early reflections arrived at the listener within the first 50 ms after the direct sound. Two conditions of the early reflections were used: one in which the early reflections increased the long-term averaged speech level by 3 dB, and the other by 6 dB. Some sound fields included only a direct sound component (D cases); others included a direct sound and early reflections ($D+E$ cases). The overall amplitudes of each of the two component groups (i.e., direct sound or early reflections) were varied, but the arrival times and relative amplitudes of individual early reflections were not changed and such details are not considered in these experiments.

Each loudspeaker also reproduced simulated ambient noise with a spectrum shape corresponding to that of an NCB40 contour.[12] The combined level of the noise signals from all seven loudspeakers was measured at the listener to be 48.4 dBA. A second noise signal with the same spectrum shape but with an overall level of 45.0 dBA was also used in the experiment.

In the first experiment, subjects were exposed to only one noise level during each of two experimental sessions. The noise signals to each loudspeaker included varied time delays so that they were not exactly coherent. The speech and noise levels used in the first experiment are summarized in Table I. All speech and noise levels were obtained with an omnidirectional measurement microphone located at the center of the listener's head position (but without the listener present). This procedure was used because it can be related to typical measurements in rooms.

2. Subjects, speech material, and procedure for the listening test

Subjects varied from 22 to 58 years of age, with nine males and five females, and they reported no hearing disabilities. Although they all spoke English everyday, for seven of the subjects English was their first language, but for the other seven subjects English was not their first language.

The speech material used in the experiments was from the Fairbanks rhyme test as modified by Latham[13] and as

TABLE I. Summary of measured acoustical quantities for *D* (direct only) cases and *D*+*E* (direct+early reflections) cases.

Direct speech level, dBA	Early speech level, dBA	Noise level, dBA	Total speech level, dBA	Total S/N, dB	Direct speech level, dBA	Early speech level, dBA	Noise level, dBA	Total speech level, dBA	Total S/N, dB
		D cases					*D* cases		
42.0	⋯	45.0	42.0	−3.0	42.0	⋯	48.4	42.0	−6.4
47.2	⋯	45.0	47.2	2.2	47.2	⋯	48.4	47.2	−1.2
51.2	⋯	45.0	51.2	6.2	51.2	⋯	48.4	51.2	2.8
53.9	⋯	45.0	53.9	8.9	53.9	⋯	48.4	53.9	5.5
57.1	⋯	45.0	57.1	12.1	57.1	⋯	48.4	57.1	8.7
62.1	⋯	45.0	62.1	17.1	62.1	⋯	48.4	62.1	13.7
		D+*E* cases					*D*+*E* cases		
42.0	42.4	45.0	45.2	0.2	42.0	42.4	48.4	45.2	−3.2
42.0	47.3	45.0	48.4	3.4	42.0	47.3	48.4	48.4	0.0
47.2	47.2	45.0	50.2	5.2	47.2	47.2	48.4	50.2	1.8
47.2	52.3	45.0	53.5	8.5	47.2	52.3	48.4	53.5	5.1
51.2	51.2	45.0	54.2	9.2	51.2	51.2	48.4	54.2	5.8
51.2	55.9	45.0	57.2	12.2	51.2	55.9	48.4	57.2	8.8

used in previous tests.[11] The test words were embedded in the sentence ‘‘Word number_ is _, write that down’’ and were spoken by a male talker. The original test lists consisted of 5 lists of 50 words, phonetically balanced within each list. In this test, each of the 250 words was used separately. Each subject listened to a total of 288 words for the combinations of 12 reflection conditions, by 2 noise levels and 12 repeats of each of these combinations. For each subject, 38 words were used twice and the repeated words were picked randomly from the 250 words. Each word was presented in a random order for each subject, and the test conditions were also presented in random order. Subjects had a break every 24 words. Each sentence was presented in a 3-s interval to the subject.

Subjects first performed a speech recognition test by writing down the first letter of each test word and then they rated the listening difficulty of each test sentence, including the target word, using the following four categories:

(0) Not difficult: no effort required, completely relaxed listening condition
(1) Slightly difficult: slight attention required
(2) Moderately difficult: moderate attention required
(3) Very difficult: considerable attention required

The authors used these same categories for listening difficulty rating in a previous study for Japanese speech.[2]

Listening difficulty was judged immediately after the speech recognition test component so that listening difficulty included the effect of the cognitive process of word recognition on this subjective rating.

Word recognition scores were obtained for each subject as the average score of the repeated tests for each condition. Listening difficulty was obtained for each subject as the percentage of responses that indicated some level of difficulty (i.e., not a ‘‘0’’ response) for each condition. This makes it possible to identify desirable conditions as those in which listeners have no difficulty in listening to the test speech material.[2] Both types of scores are presented as mean scores over all subjects.

B. Results of listening test to speech

1. The relation between listening difficulty and word intelligibility

The first comparisons were based on the results of tests in which subjects performed speech intelligibility tests and listening difficulty ratings for sound fields with varied speech signal-to-noise ratio (S/N) and for two types of reflection conditions. In one series of tests the sound fields consisted of only a direct sound and varied S/N was obtained by varying the amplitude of the direct speech sound in combination with either 45.0 or 48.4 dBA of constant noise. In the other series of tests, three levels of direct speech sound were used and the S/N was varied by adding early reflections to increase the total speech levels by 3 or 6 dBA relative to the direct speech levels. These were all presented in combination with the same two constant noise levels. Figure 1 and Fig. 2 show the

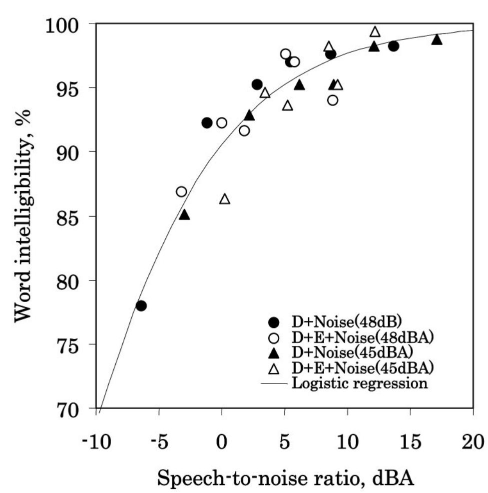

FIG. 1. Mean intelligibility scores for each sound field condition for all listeners for *D* (direct only) cases (filled symbols) and *D*+*E* (direct +early reflections) cases (open symbols) in 48-dBA noise (circles) and 45-dBA noise (triangles). Logistic regression curve for all data is also presented.

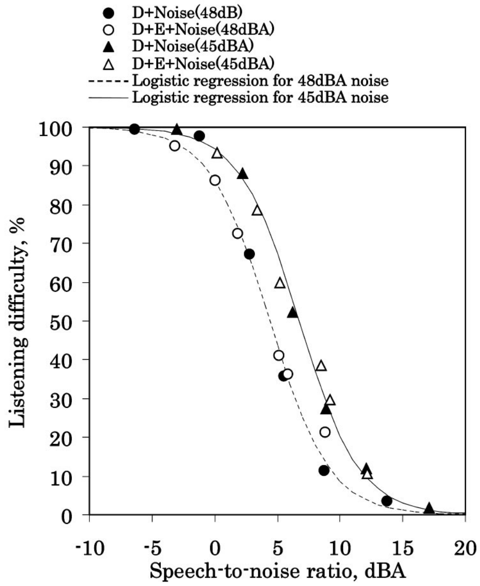

FIG. 2. Mean listening difficulty ratings of each sound-field condition for D (direct only) cases (filled symbols) and $D+E$ (direct+early reflections) cases (open symbols) in 48-dBA noise (circles) and 45-dBA noise (triangles). Logistic regression curves for all data with 45-dBA noise (solid line) and of 48-dBA noise (dashed line) are also presented.

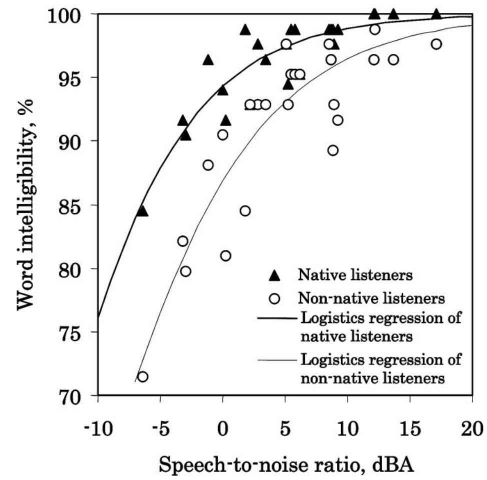

FIG. 3. Mean intelligibility scores of native English speakers (gray triangle) and non-native English speakers (open circle) for all sound-field conditions. Logistic regression curves are included for the native English speakers (solid line) and for the non-native English speakers (dashed line).

resulting plots of word recognition scores and listening difficulty ratings versus S/N, respectively. Both ratings show strong relationships with S/N.

Although word recognition scores exceed 90% for S/N greater than 0 dBA, listening difficulty is about 90% at a S/N of 0 dBA and shows the greatest variation for S/N above 0 dBA. Listening difficulty decreases monotonically as S/N varies from -2.5 to 15 dBA. In other words, listening difficulty varies over a range of 97% (i.e., from 2% to 99%) for the experimental conditions. However, word recognition scores varied over a range of only 21% (i.e., from 78% to 99%) and are seen to be a less sensitive rating of these acoustical conditions.

Analysis of variance (ANOVA) was employed to compare the sensitivity of word recognition and listening difficulty scores. The experimental conditions and individual differences were the two factors tested by the repeated measures ANOVA. The result of ANOVA for word recognition scores indicated that there was a significant effect of conditions ($p<0.0001$) and there was also a significant effect of individual subject differences ($p<0.0001$). Tukey's honestly significant difference (HSD) test for multiple comparisons[14] was employed, and it indicated that differences between conditions of more than 9.12% would be significant. A difference of 9.12% is 42.6% of the complete range of the speech recognition scores.

The ANOVA results for the listening difficulty ratings indicate that there are significant effects of varied condition ($p<0.0001$) and also of the differences among subjects ($p<0.0001$). HSD is 22.8% ($p<0.05$), which is 23.4% of the complete range of listening difficulty ratings, and this is about half of that for word intelligibility. This HSD shows that the listening difficulty ratings have approximately double the sensitivity of the word intelligibility scores.

There was no pair of conditions, with and without early reflections and both having almost the same speech-to-noise ratio, which had difficulty ratings different by 23.4% or more. This indicates that, for listening difficulty ratings, adding early reflections is equivalent to increasing the energy of the direct sound to the same total speech level. In other words, early reflections are equivalent to increased direct speech energy in these conditions without reverberant speech sound.

2. Individual differences among subjects

Because significant differences among individuals were found, ANOVA was employed to test for differences between the subject groups of native and non-native English speakers, all of whom used English every day. Each group had seven subjects. The experimental conditions and subject groups were the two factors included in the ANOVA analysis. The interaction between conditions and subject groups was not significant either for word recognition score or for listening difficulty.

The result of the ANOVA for word recognition scores indicated that there was a significant difference between the two subject groups ($p<0.0001$). Figure 3 shows that there are clearly different trends with respect to S/N for the word intelligibility scores for each subject group. The regression curves in Fig. 3 indicate that the difference in S/N between the two subject groups was 5.7 dB for a 90% word intelligibility score. This means that non-native English speakers had a 5.7-dB disadvantage relative to native English speakers in terms of speech recognition scores. This is comparable to results by Buus *et al.*,[15] who found non-native listeners to require between 3- and 12-dB better conditions depending on their amount of experience with the language of the test. Somewhat similar results were reported by Nabelek and Donahue[16] for varied reverberation time and by van Wijngaarden *et al.*[17] for Dutch subjects listening to both English and German speech material.

On the other hand, for listening difficulty ratings, the ANOVA analysis indicated that there was not a significant difference between the two subject groups ($p > 0.05$).

One can conclude that although word intelligibility and listening difficulty ratings vary among subjects, listening difficulty is not affected by the mother tongue of the subjects. Thus, listeners of varied language skills can be used to provide similar listening difficulty ratings of conditions for speech. However, these ratings would not reflect the expected lower intelligibility scores from listeners less familiar with the language of the test.

3. Effect of noise level on listening difficulty

The difference between the two logistic regression curves (where difficulty is 50%) in Fig. 2 indicates that the effective S/N difference between the two sets of results was only 1.7 dBA. This 1.7-dB difference is only half of the actual difference of noise levels and would correspond to a 20%–30% difference in listening difficulty ratings. This difference is approximately the same as the HSD ($p < 0.05$) required to indicate a significant difference between a pair of conditions for listening difficulty. It is likely that this difference is due to context effects in the experiment. Although it was hoped that subjects would judge listening difficulty absolutely, they may have tended to rate conditions relative to the complete range of conditions to which they were exposed in each part of the experiment. In particular, the cases for each noise level were presented as two subgroups in which only one ambient noise level was experienced in each subgroup. This led to a slightly different range of S/N for the two parts of the experiment, and is probably the cause of the 1.7-dB shift between the two sets of listening difficulty rating results.

ANOVA was employed to test the significance of the difference between the results for ambient noise of 48.4 dBA and those of 45.0 dBA. Noise level, speech level, and the interaction between noise level and speech level were included as factors in the ANOVA. The result of the ANOVA was that there was only a marginally significant difference between the two ambient noise conditions ($p < 0.1$).

Even though the difference was not statistically significant at $p < 0.05$, Fig. 2 suggested that possible context effects should be minimized in subsequent experiments. Accordingly, the second experiment, discussed in the next section, included a wider range of conditions and they were all presented in a single experimental listening session.

III. LISTENING TESTS FOR SPEECH IN AMBIENT NOISE AND REVERBERATION TO CONFIRM THE BENEFIT OF EARLY REFLECTIONS

A. Method

1. Sound-field simulation procedures

All simulated sound fields were produced using a seven-channel electro-acoustic system with loudspeakers arranged around the listener in an anechoic room similar to that described for the first experiments. The seven loudspeakers were located at a distance of 1.6 m from the listener. The delayed early reflections and reverberation were created using programmable digital signal processors (Yamaha DME32) that could be changed under computer control via a MIDI interface. The loudspeaker responses were corrected to be flat ± 1 dB from 80 Hz to 12 kHz.

The loudspeaker located directly in front of the listener produced the simulated direct sound and the other six loudspeakers each produced one early reflection of the speech sounds. The early reflections arrived at the listener with varied delays to distribute them in a realistic manner over the first 50 ms after the direct sound. The details of individual reflections were not varied; only the overall amplitude of the six early reflections was varied as a group. Two levels of the early reflections were used in the experiment, one of which increased the long-term averaged speech level relative to that of the direct sound by 3 dB, and the other by 6 dB. Reverberant speech was produced via all seven loudspeakers with slightly different levels and delay times for each speaker to create a diffuse impression for the reverberant speech component.

Each loudspeaker also reproduced simulated ambient noise with a spectrum shape corresponding to that of an NCB40 contour and with a measured overall level at the position of the listener of 48.6 dBA. The noise signals to each loudspeaker were not exactly coherent to minimize interference effects at the listener and to create the impression of a diffuse sound field for the simulated ambient noise.

There were three series of conditions, which are described in Table II. In one series the sound fields consisted of only a direct sound (D cases) and varied S/N was obtained by varying the amplitude of the direct speech sound relative to the constant level of ambient noise. In the second series, the sound fields consisted of a direct sound and two levels of reverberation ($D+$Rev.A or $D+$Rev.B cases). The reverberation time was 1.1 s for both reverberant cases, but the ratio of early to late arriving speech sound (C50) varied as described in Table II. The reverberant speech level was 51.6 dBA for the more reverberant case called "Rev.A" and 45.8 dBA for the less reverberant case called "Rev.B." There were four levels of direct sound for each reverberant case, increasing in 3-dBA steps from 49 dBA. In the third series ($D+E+$Rev.A or $D+E+$Rev.B cases), two levels of early reflections, which increased the effective signal level by 3 and 6 dBA, were added to the 49-dBA direct sound level condition with each of "Rev.A" and "Rev.B" late-arriving energy and were compared with cases which had the same effective signal level. The overall amplitudes of each of the three component groups (direct sound, early reflections, and reverberant sound) were varied but the arrival times and relative amplitudes of individual early reflections were not changed.

The measured levels of each component group are summarized in Table II for each of the test conditions. The levels of the direct, early, and reverberant speech sounds are listed where appropriate. This table also lists the effective speech levels consisting of the sum of the direct sound and the early reflection energy arriving within 50 ms after the direct sound, along with the corresponding effective signal-to-noise ratios (E-S/N).

To avoid context effects for a limited series of conditions, as discussed in the previous section, subjects experi-

TABLE II. Summary of measured acoustical quantities for D (direct only) cases, $D+\text{Rev}$ (direct+reverberant sound) cases and $D+E+\text{Rev.}$ (direct+early reflections+reverberant sound) cases. The ambient noise level was fixed at 48.6 dBA for all cases listed in this table.

Direct speech level, dBA	Early speech level, dBA	Reverb speech level, dBA	Effective speech level, dBA	Total speech level, dBA	Effective S/N, dB	Total S/N, dB	C50, dBA	U50, dBA	STIr (male)	RT (0.5–1 kHz), s
					D cases					
42.7	⋯	⋯	42.7	42.7	−5.9	−5.9	⋯	−5.9	0.29	⋯
45.7	⋯	⋯	45.7	45.7	−2.9	−2.9	⋯	−2.9	0.39	⋯
48.7	⋯	⋯	48.7	48.7	0.1	0.1	⋯	0.1	0.49	⋯
51.7	⋯	⋯	51.7	51.7	3.1	3.1	⋯	3.1	0.59	⋯
54.7	⋯	⋯	54.7	54.7	6.1	6.1	⋯	6.1	0.68	⋯
57.7	⋯	⋯	57.7	57.7	9.1	9.1	⋯	9.1	0.78	⋯
60.7	⋯	⋯	60.7	60.7	12.1	12.1	⋯	12.1	0.88	⋯
63.7	⋯	⋯	63.7	63.7	15.1	15.1	⋯	15.1	0.94	⋯
					D+Rev. cases					
49.4	⋯	51.6	49.4	53.6	0.8	5.0	−2.1	−3.8	0.38	1.1
52.4	⋯	51.6	52.4	55.0	3.8	6.4	0.8	−0.9	0.45	1.1
55.3	⋯	51.6	55.3	56.9	6.7	8.3	3.8	2.1	0.54	1.1
58.3	⋯	51.6	58.3	59.2	9.7	10.6	6.8	5.1	0.64	1.1
49.3	⋯	45.8	49.3	50.9	0.7	2.3	3.7	−0.2	0.43	1.1
52.3	⋯	45.8	52.3	53.2	3.7	4.6	6.7	2.7	0.53	1.1
55.3	⋯	45.8	55.3	55.8	6.7	7.2	9.7	5.7	0.62	1.1
58.3	⋯	45.8	58.3	58.6	9.7	10.0	12.7	8.7	0.72	1.1
					D+*E*+Rev. cases					
49.3	49.4	51.6	52.4	55.0	3.8	6.4	0.8	−0.9	0.44	1.1
49.3	54.0	51.6	55.3	56.8	6.7	8.2	3.7	1.8	0.51	1.1
49.3	49.4	45.8	52.4	53.2	3.8	4.6	6.7	2.7	0.50	1.1
49.3	53.9	45.8	55.2	55.7	6.6	7.1	9.6	5.4	0.58	1.1

enced a full range of experimental conditions in one single test. That is, cases varied from near-zero listening difficulty (S/N of 15 dBA) to nearly complete (100%) listening difficulty (S/N of −6 dBA).

2. Subjects, speech material, and procedure for the listening test

Eleven male subjects and two female subjects were used for the experiment. Subjects varied from 21 to 58 years of age and they didn't report any hearing disabilities. Two non-native English speakers were used in this experiment but their scores on the word intelligibility test were not included in the results.

The speech material used in this experiment was the same as in the previous experiment. Each subject listened to 10 sentences for each of 20 conditions and they heard each test sentence only once. Test words and test conditions were presented in random order to each subject. Subjects could have a small break every 25 words. Each sentence was presented after the subject responded to the former sentence. Hence, the rate of presentation depended on the subject.

The experimental procedure was almost the same as in the previous experiment except a small laptop PC with a 6 in. screen and a full-sized keyboard, positioned on the knees of the subject, was used instead of pen and paper. Subjects were asked to use the keyboard to give the missing first letter of each test word in the speech recognition part of the test. After listening to each sentence, they also rated the listening difficulty by typing the number for the category of difficulty.

To minimize individual differences in the results, subjects were informed about the procedure of the experiment and the definition of listening difficulty in a training session. In this practice session, subjects listened to more than the full range of experimental conditions (S/N was systematically varied in 3-dB steps from +18 to −9 dB and then in reverse order from −9 to +18 dB varying only the direct sound and ambient noise). They responded to the word recognition test and gave a listening difficulty rating to get them used to the idea of listening difficulty and to the complete range of conditions. All of the subjects switched their responses from "not-difficult" to "difficult" and vice versa in the middle of the range of S/N in this training session.

B. Results of the listening tests

1. ANOVA for comparing sensitivity of word intelligibility and listening difficulty

The result of the ANOVA for word recognition scores (the two non-native English speakers were removed) showed that there was a significant effect of varied condition ($p < 0.0001$) but there was not a significant effect of individual subject differences ($p > 0.05$). Tukey's honestly significant difference (HSD) test was employed for multiple comparisons. HSD was calculated to be 6.88% ($p < 0.05$), which is 42.0% of total range of the scores.

The result of the ANOVA for listening difficulty indicated that there was a significant effect of varied condition ($p < 0.0001$). There was also a significant effect of differences among individual subjects ($p < 0.0001$). HSD for listening difficulty was 24.8%($p < 0.05$), which is 25.1% of the complete range of listening difficulty.

Although there is a significant effect of the differences among individual subjects, the HSD is a smaller portion of the complete range of listening difficulty scores than is the HSD for the word intelligibility scores. This is true even

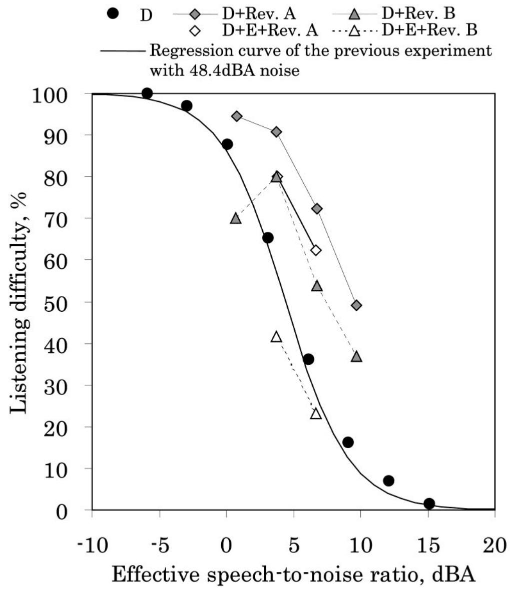

FIG. 4. Relation between effective speech-to-noise ratio and mean difficulty rating of D (direct sound) cases (filled circle), $D+$Rev. (direct sound $+$reverberant sound) cases (gray symbols), and $D+E+$Rev. (direct sound $+$early reflections$+$reverberant sound) cases (open symbols). The logistic regression curve for the 48.4-dBA-noise case of the previous experiment is also presented.

though there is not a significant effect on word recognition scores of individual differences among subjects when the mother tongue effect was removed.

This again shows that listening difficulty can better discriminate among these experimental conditions than can word recognition scores.

2. Comparison of listening difficulty results with those in the previous experiment

Figure 4 presents the relation between listening difficulty ratings and the effective speech-to-noise ratio (E-S/N), in which the energy of direct sound and the early reflections within 50 ms of direct sound are summed as the effective speech energy. Figure 4 also includes the regression curve obtained from the results of conditions with 48.4-dBA noise in the previous experiment.

The cases with only a direct sound component (filled circles in Fig. 4) are close to the regression curve from the former experiment. Additionally, listing difficulty ranged over almost the full scale from 1.5% to 100% as intended to avoid context effects. This result demonstrates the repeatability of the listening difficulty measure, at least for the equivalent conditions in the previous experiment.

3. Benefit of early reflections in noise and reverberation for listening difficulty

The series of conditions in this experiment were created to confirm that the effect of early reflections also exists in more realistic cases that also included later-arriving speech sounds (reverberation).

The results in Fig. 4 show that the reverberant level influences listening difficulty ratings. The more reverberant Rev.*A* cases created more difficult listening conditions than the less reverberant Rev.*B* cases at the same E-S/N.

The lowest S/N (0.7 dB) case for Rev.*B* without early reflections (filled triangles) deviates from the overall trend of these results and showed lower difficulty for this S/N value than expected. The difference between the lowest case and the second lowest case is smaller than the HSD and is not statistically significant ($p>0.05$). In a separate test, five subjects compared these two conditions ten times and 90% of their responses were the opposite of the result in the main experiment as was expected. All of the subjects who participated in this trial reported that there was a subtle difference between the two conditions and it might relate to a difference in loudness. In all other cases, adding early reflections increased the effective S/N and decreased the resulting listening difficulty rating. This one unusual point was thought to be due to the large scatter in listening difficulty scores; this was verified in a paired comparison test described in next section.

From Fig. 4, it is seen, that for sound fields including Rev.*A* with the direct sound (filled diamonds), the listening difficulty ratings are about the same as those for the direct sound only cases having a 5-dB lower S/N. Adding the reverberant speech level from Table II (51.6 dB for Rev.*A*) to the noise level (48.6 dB) increases the total detrimental sound level by about 5 dB relative to the noise alone. This 5-dB increase in detrimental sound level would relate to a 5-dB decrease in useful-to-detrimental sound ratios between the direct-sound-only cases and the direct-plus-reverberant cases, and supports the use of useful-to-detrimental ratio concept (U50) (Refs. 18, 19) to explain the effects of added reverberant speech sound on listening difficulty for these cases.

Listening difficulty ratings for cases with early reflections are much lower than those for cases without early reflections at the same E-S/N. Early reflections decrease listening difficulty by 10% for Rev.*A* [which is less than the HSD ($\mathrm{HSD}=24.8\%, p<0.05$)] and by 40% for Rev.*B* (which is greater than HSD) compared to other cases with the same E-S/N. This may be due to early reflections effectively increasing the time window for integrated useful early energy and thus including the early part of late-arriving sound (more than 50 ms after the direct sound) in the integrated useful energy.

Figure 5 presents the relation between listening difficulty ratings and U50(A) values (A-weighted useful-to-detrimental ratio with 50-ms early time interval). As expected, similar trends versus U50(A) are seen for the only-direct-sound cases and the direct-sound-plus-reverberation cases. The cases with early reflections deviate from the main trend and show lower difficulty for a particular U50(A) value than for the cases without early reflections. STIr(Male), which is a new version of the Speech Transmission Index for male voices,[20] shows almost the same relation with listening difficulty as found for U50(A) in Fig. 5. These results may indicate that these measures could be adjusted to better reflect the benefits of early reflections on the listening difficulty scores. Although the previous study[11] discussed the benefit of early reflections on speech recognition scores rang-

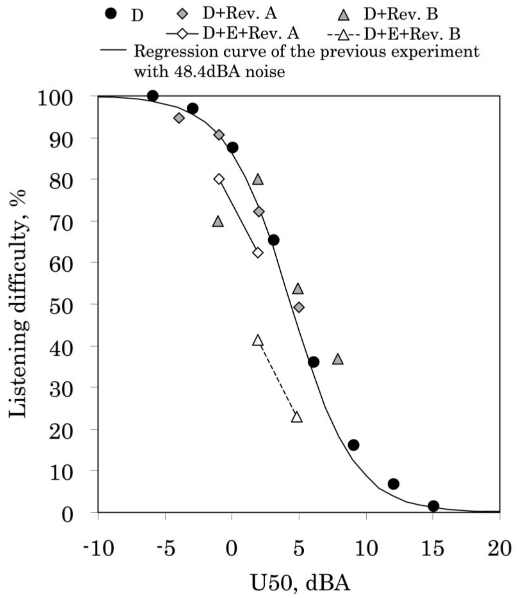

FIG. 5. Relation between A-weighted useful-to-detrimental ratio [U50(A)] and mean difficulty rating of *D* (direct sound) cases (filled circle), *D* + Rev. (direct sound + reverberant sound) cases (gray symbols), and *D* + *E* + Rev. (direct sound + early reflections + reverberant sound) cases (open symbols). The logistic regression curve of the 48.4-dBA-noise case of the former experiment is also presented.

ing from 85% to 100% for normal-hearing listeners, listening difficulty better discriminates among conditions in rooms to illustrate the effect of early reflections, as this study shows.

Because the scatter in the listening difficulty ratings, due to individual differences among subjects, is large, it is difficult to precisely relate listening difficulty ratings with physical indices. More precise ratings for some of the cases used in this experiment were obtained using a paired comparison test presented in the next section.

IV. SCHEFFE'S PAIRED COMPARISON TEST TO CONFIRM THE BENEFIT OF EARLY REFLECTIONS ON LISTENING DIFFICULTY TO SPEECH

In order to show the significant benefit of early reflections to listening difficulty, and to discuss the relation of listening difficulty scores with physical numbers, Scheffe's method of paired comparison test[21] was used.

A. Method

1. Sound fields

Seven of the conditions used in the previous experiment were used for the paired comparison tests. The electroacoustic system and the anechoic chamber used in this experiment were the same as in the previous listening test. Two direct-sound-only-plus-noise cases (D), three cases with direct sound plus Rev.B (D+Rev.B), and two cases with direct sound, reverberant sound (Rev.B), and early reflections ($D+E+$Rev.B) were used. All conditions are summarized in Table III.

2. Subjects and procedure

Eleven of the subjects who participated in the former experiment were used. Subjects were asked to rate the differences for each pair in one of five categories. They did this by typing a number on the keyboard with the following descriptions presented on the screen of the tiny PC, located in front of subjects, and after listening to each pair of sentences:

(1) Former is much more difficult than latter.
(2) Former is more difficult than latter.
(3) Former is as difficult as latter.
(4) Latter is more difficult than former.
(5) Latter is much more difficult than former.

The categories were assigned scores of −2, −1, 0, 1, and 2 corresponding to the first to the fifth responses in the list above, respectively. A total of 42 different pairs of speech conditions was presented twice to each subject. (Because this psychological scale is a relative scale, it could have used the word "easy" instead of "difficult.")

B. Results

1. Confirmation of significant benefit of early reflections to listening difficulty

The result of the ANOVA of the results from the psychological scale of listening difficulty showed that there was a significant effect of test condition ($p<0.0001$) and there was not a significant effect of the differences among individual subjects ($p>0.05$). The Yardstick[21] (same concept as HSD) was calculated to be 0.13 ($p<0.05$).

TABLE III. Summary of measured acoustical quantities for Scheffe's paired comparison test selected from some of cases in the former experiment in Table II. The ambient noise level was fixed at 48.6 dBA for all cases listed in this table.

Condition	Direct speech level, dBA	Early speech level, dBA	Reverb speech level, dBA	Effective speech level, dBA	Total speech level, dBA	Effective S/N, dB	Total S/N, dB	C50, dBA	U50, dBA	STIr (male)	RT (0.5–1 kHz), s
					D cases						
(i)	48.7	⋯	⋯	48.7	48.7	0.1	0.1	⋯	0.1	0.49	⋯
(ii)	57.7	⋯	⋯	57.7	57.7	9.1	9.1	⋯	9.1	0.78	⋯
					D+Rev*B* cases						
(iii)	49.3	⋯	45.8	49.3	50.9	0.7	2.3	3.7	−0.2	0.43	1.1
(iv)	52.3	⋯	45.8	52.3	53.2	3.7	4.6	6.7	2.7	0.53	1.1
(v)	55.3	⋯	45.8	55.3	55.8	6.7	7.2	9.7	5.7	0.62	1.1
					D+*E*+Rev*B* cases						
(vi)	49.3	49.4	45.8	52.4	53.2	3.8	4.6	6.7	2.7	0.50	1.1
(vii)	49.3	53.9	45.8	55.2	55.7	6.6	7.1	9.6	5.4	0.58	1.1

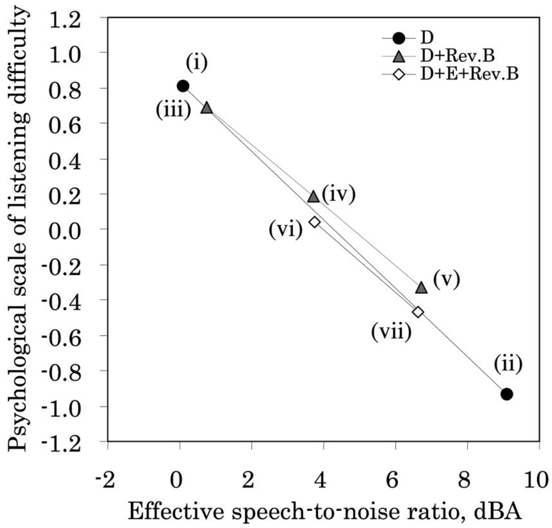

FIG. 6. Relation between effective speech-to-noise ratio and psychological scale value of listening difficulty of *D* (direct sound) cases (filled circle), *D*+Rev.*B* (direct sound+reverberant sound Rev.*B*) cases (gray triangle), and *D*+*E*+Rev.*B* (direct sound+early reflections+reverberant sound Rev.*B*) cases (open diamond). The data points are numbered from (i) to (vii) as defined in Table III.

Comparing the conditions with direct sound and those with direct sound plus reverberation, listening difficulty was increased by reverberation as shown in Fig. 6. When early reflections were added, listening difficulty was reduced relative to cases without early reflections at the same E-S/N. Conditions (i)–(vii) are described in Table III. If the difference between two conditions is greater than the Yardstick, it indicates that these two conditions are significantly different. The difference between condition (iv) (without early reflections) and condition (vi) (with early reflections) is greater than the Yardstick, and the difference between condition (v) (without early reflections) and condition (vii) (with early reflections) is also greater than the Yardstick. These results suggest that the listening difficulty of the conditions with early reflections (vi) is less than the condition without early reflections (iv), which has the same effective speech energy as condition (vi). The same could be said for the relation between (v) and (vii). This result significantly demonstrates that early reflections increase the effectiveness of the speech sounds more than expected due to the summation of the direct and early reflection energy in cases with noise and reverberation.

The key finding from Fig. 5 is the result that the conditions with early reflections and reverberation were rated as less difficult than the direct-sound-only case having the same E-S/N.

Adding reverberant speech (*D*+Rev.*B* cases) to the direct-sound-only cases (*D*) caused increases in listening difficulty that increased with increasing E-S/N. This result suggests that the reason for the lowest E-S/N (0.7 dB) case for Rev.*B* without early reflections in the previous experiment deviating from the overall trend of results was due to the scatter of listening difficulty ratings.

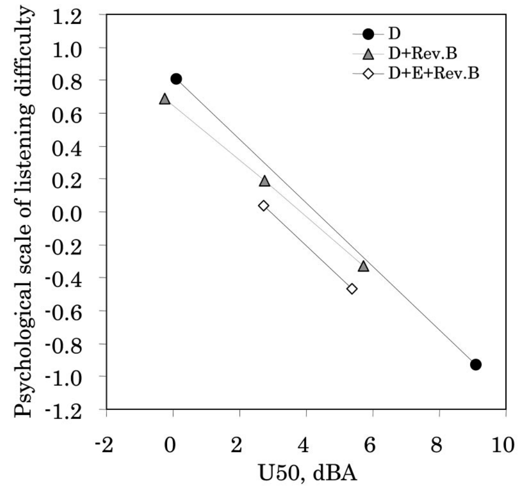

FIG. 7. Relation between A-weighted useful-to-detrimental ratio [U50(A)] and psychological scale values of listening difficulty for *D* (direct sound) cases (filled circle), *D*+Rev.*B* (direct sound+reverberant sound Rev.*B*) cases (filled triangle), and *D*+*E*+Rev.*B* (direct sound+early reflections +reverberant sound Rev.*B*) cases (open diamond).

2. Relation between listening difficulty and physical measures

As Fig. 6 shows, listening difficulty ratings are well related to E-S/N for the *D* cases and the *D*+*E*+Rev.*B* cases. The *D*+Rev.*B* cases are expected to differ from the main trend as E-S/N increases. This is because E-S/N does not take into account the detrimental effect of reverberation as the small differences in listening difficulty show.

Figure 7 shows the relation between the psychological scale of listening difficulty and U50(A). These results suggest that U50(A) overestimates the detrimental effects of reverberant sound in these results. If the time interval for useful energy is increased to 170 ms for *D*+Rev.*B* cases [(iii), (iv), and (v) in Table III], and to 240 ms for *D*+*E*+Rev.*B* cases [(vi) and (vii)], the variation of ratings with useful-to-detrimental ratios would agree with the direct-sound-only cases [(i) and (ii)] as illustrated in Fig. 8. This suggests that the effective early time interval may vary due to differences in the addition of early reflection components, and that a simple energy addition over a fixed 50-ms early time interval may not always be appropriate.

STIr values show the same trend as do the U50(A) results and similarly overestimate the detrimental effects of reverberant sound seen in Fig. 9.

V. DISCUSSION

Figure 10 shows the variation of word intelligibility scores of native English speakers and listening difficulty ratings with U50(A), for all of the conditions in this study. The listening difficulty ratings do exhibit a reasonable amount of scatter. This may be partially due to systematic errors in U50 values that were based on a fixed 50-ms early time interval as pointed out in the results of this study. Further efforts are required to find a more appropriate procedure for determining the boundary for the early time interval and to better understand the process by which early reflection energy is integrated into the useful speech energy in complete room

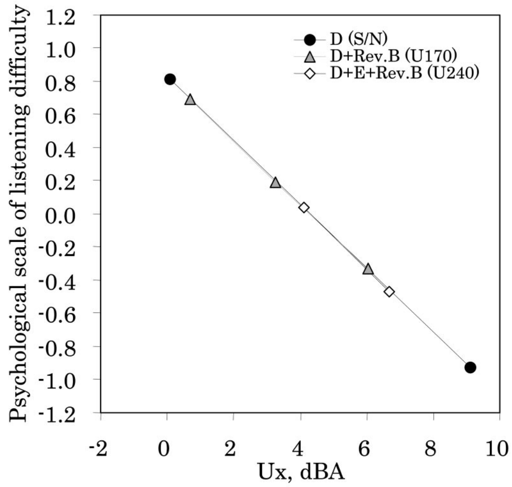

FIG. 8. Relation between A-weighted useful-to-detrimental ratio [Ux(A)] and psychological scale value of listening difficulty for D (direct sound) cases (filled circle), $D+\text{Rev.}B$ (direct sound + reverberant sound Rev.B) cases (filled triangle, $x=170$ ms), and $D+E+\text{Rev.}B$ (direct sound + early reflections + reverberant sound Rev.B) cases (open diamond, $x=240$ ms).

impulse responses. Several previous studies[9,22–24] give some clues as to how one might improve the method of evaluating the effect of room acoustics on speech communication.

This study clearly shows, that early reflections at least have an effect equivalent to amplifying the direct sound by as much as the energy increase they provide, and they improve the E-S/N for listening difficulty under noisy conditions. The results of experiments in noisy and reverberant sound fields suggest that early reflections tend to expand the time window of the useful early speech energy. This finding suggests that some late-arriving sound (i.e., greater than 50 ms after the direct sound) is helpful, and that this would influence the determination of optimum reverberation time criteria for speech communication in noise.

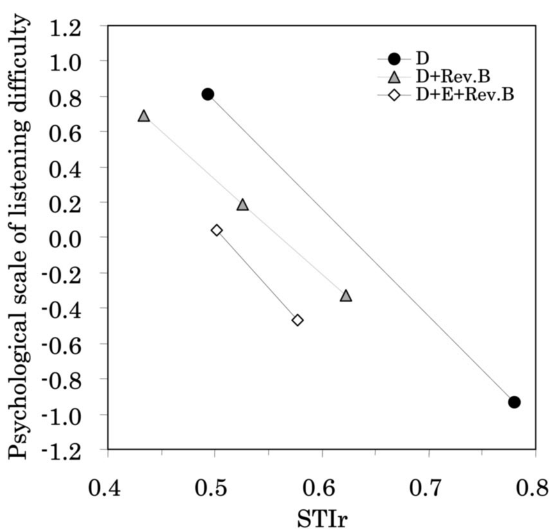

FIG. 9. Relation between revised STI for male (STIr) and psychological scale values of listening difficulty for D (direct sound) cases (filled circle), $D+\text{Rev.}B$ (direct sound + reverberant sound Rev.B) cases (filled triangle), and $D+E+\text{Rev.}B$ (direct sound + early reflections + reverberant sound Rev.B) cases (open diamond).

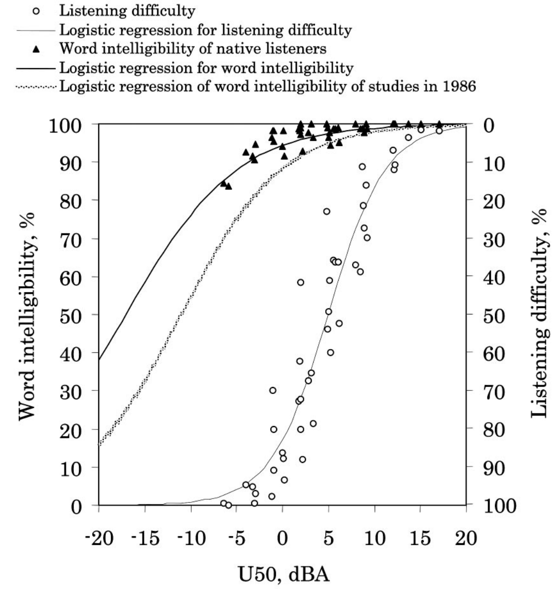

FIG. 10. Relation of word intelligibility (filled triangle) and listening difficulty (open circle) for all conditions in this study with the A-weighted useful-to-detrimental ratio [U50(A)]. Also shown is the regression curve to earlier results for word recognition scores in rooms.

The experimental conditions in this study ranged from -6 to $+15$ dBA in terms of S/N, and this range covered almost all conditions we might find in normal public spaces as identified by Pearsons *et al.*[1] The new results showed that listening difficulty ranged from 0% to 100% in this range of conditions. Values of 5% and 95% of listening difficulty correspond to S/N values of -4.5 and 14.5 dBA, respectively. On the other hand, for this same range of S/N values, word intelligibility scores only varied from 90% to 100%. This small range of word recognition scores makes it more difficult to consider the influence of early reflections on speech communication using word intelligibility scores.

In a previous study,[2] the authors discussed the range of sound-field conditions that would be most suitably evaluated using intelligibility scores, difficulty ratings, or sound quality and easiness ratings. Each rating is most appropriate for a particular range of conditions. Figure 10 also presents the regression line of intelligibility scores of old studies using the same rhyme test and measured in actual sound fields (classroom, gym, auditorium, etc.).[25,26] The curve is slightly lower than the new data presented in this study, but still indicates 90% intelligibility at a U50 of 0 dBA. (In the older studies U50 values were probably a few dB too high, because speech levels were estimated from the known output of the source using simple diffuse field theory.) The new results in the current paper clearly illustrate the range of S/N or U50 for which word intelligibility scores are more appropriate (up to 0 dBA) and the range for which listening difficulty ratings are more appropriate (-4.5 to 14.5 dBA).

One problem with listening difficulty ratings is the scatter among the results of the different subjects, even though they had received a training session. One solution to this problem is to use paired comparison tests, but for large sets of conditions, paired comparison tests can be very time consuming. More effort is required to minimize scatter of the difficulty ratings.

VI. CONCLUSIONS

The results demonstrate that

(1) Listening difficulty ratings better discriminate among conditions commonly found in spaces intended for speech communication than do word recognition scores.
(2) For S/N values ranging from −4.5 to +14.5 dBA, listening difficulty scores vary from 5% to 95%, a range of 90%. On the other hand, word intelligibility scores vary only a small amount over the upper part of this range, corresponding to near to acceptable conditions for speech communication. Word intelligibility scores are about 95% at an S/N of +1 dBA and start to decrease increasingly rapidly below this S/N value.
(3) Early reflection energy has at least the same effect on speech intelligibility and listening difficulty ratings as an equivalent increase in direct sound level. When reverberation exists, the benefit of early reflections is more than expected due to the added early reflection energy.
(4) Listening difficulty isn't affected by the mother tongue of subjects in the English rhyme test. It is therefore a more widely applicable test for evaluating the quality of conditions for speech communication. However, listening difficulty ratings would not reflect the decreased intelligibility scores of non-native listeners.

ACKNOWLEDGMENTS

This study was done while the first author was on sabbatical leave from Faculty of Engineering, Tohoku University, Japan, granted by The Kajima Foundation. The research project was partially supported by the Twenty-first Century Center of Excellence (COE) Program, ''Design Strategy towards Safety and Symbiosis of Urban Space'' awarded to Graduate School of Science and Technology, Kobe University. The Ministry of Education, Culture, Sports, Science and Technology of Japan sponsored the Program.

[1] K. S. Pearsons, R. L. Bennett, and S. Fidell, ''Speech levels in various noise environments,'' Report No. EPA-600/1-77-025, Washington, D.C. (1977).
[2] M. Morimoto, H. Sato, and M. Kobayashi, ''Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces,'' J. Acoust. Soc. Am. **116**, 1607–1613 (2004).
[3] H. Sato, H. Yoshino, and M. Nagatomo, ''Relationship between speech transmission index and easiness of speech perception in reverberatory fields,'' J. Acoust. Soc. Am. **103**, Pt. 2, 2999 (1998).
[4] J. E. Greenberg, J. G. Desloge, and P. M. Zurek, ''Evaluation of array-processing algorithms for a headband hearing aid,'' J. Acoust. Soc. Am. **113**, 1646–1657 (2003).
[5] J. F. Feuerstein, ''Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort,'' Ear Hear. **13**, 80–86 (1992).
[6] F. Apoux, O. Crouzet, and C. Lorenzi, ''Temporal envelope expansion of speech in noise for normal-hearing and hearing impaired listeners: Effects on identification performance and response times,'' Hear. Res. **153**, 123–131 (2001).
[7] D. W. Downs, ''Effects of hearing aid use on speech discrimination and listening effort,'' J. Speech Hear Disord. **47**, 189–193 (1982).
[8] C. B. Hicks and A. M. Tharpe, ''Listening effort and fatigue in school-age children with and without hearing loss,'' J. Speech Lang. Hear. Res. **45**, 573–584 (2002).
[9] J. P. A. Lochner and J. F. Burger, ''The influence of reflections on auditorium acoustics,'' J. Sound Vib. **1**, 426–454 (1964).
[10] ANSI S12.60-2002, ''Acoustical performance criteria, design requirements and guidelines for schools,'' (American National Standards Institute, New York, 2002).
[11] J. S. Bradley, H. Sato, and M. Picard, ''On the importance of early reflections for speech in rooms,'' J. Acoust. Soc. Am. **113**, 3233–3244 (2003).
[12] L. L. Beranek, ''Balanced noise-criterion (NCB) curves,'' J. Acoust. Soc. Am. **86**, 650–664 (1989).
[13] H. G. Latham, ''The signal-to-noise ratio for speech intelligibility—An auditorium acoustics design index,'' Appl. Acoust. **12**, 253–320 (1979).
[14] J. W. Tukey, ''The problem of multiple comparisons,'' Mimeographed Monograph (1953), appears in full in *Collected work of J. W. Tukey*, edited by H. Braun (Chapman & Hall Inc., New York, 1994), Vol. VII.
[15] S. Buus, M. Florentine, B. Scharf, and G. Canevet, ''Native, french listeners' perception of american english in noise,'' Proceedings of Inter Noise 86, Cambridge, MA, July 1986.
[16] A. K. Nabelek and A. M. Donahue, ''Perception of consonants in reverberation by native and non-native listeners,'' J. Acoust. Soc. Am. **75**, 632–634 (1984).
[17] S. J van Wijngaarden, H. J. M. Steeneken, and T. Houtgast, ''Quantifying the intelligibility of speech in noise for non-native listeners,'' J. Acoust. Soc. Am. **111**, 1906–1916 (1998).
[18] J. S. Bradley, ''Relationships among measures of speech intelligibility in rooms,'' J. Audio Eng. Soc. **46**, 396–405 (1998).
[19] J. S. Bradley, R. D. Reich, and S. G. Norcross, ''On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility,'' J. Acoust. Soc. Am. **106**, 1820–1828 (1999).
[20] H. J. M. Steeneken and T. Houtgast, ''Mutual dependency of the octave-band weights in predicting speech intelligibility,'' Speech Commun. **28**, 109–123 (1999).
[21] H. Scheffe, ''An analysis of variance for paired comparisons,'' J. Am. Stat. Assoc. **47**, 381–400 (1952).
[22] R. Plomp, ''Rate of decay of auditory sensation,'' J. Acoust. Soc. Am. **36**, 277–282 (1964).
[23] V. R. Thiele, ''Richtungsverteilung und zeitfolge der schallruckwurfe in raumen,'' Acustica **3**, 291–302 (1953). (Title in English: ''Directional distribution and time response to impulse sounds in rooms.'')
[24] Y. Tahara and H. Sato, ''Auditory integration characteristics based on critical time delay of echo perception for improvement of indices for room acoustics,'' Proceedings of International Symposium on Room Acoustics, Awaji, Japan, 8 pages on CD-ROM (2004).
[25] J. S. Bradley, ''Predictors of speech intelligibility in rooms,'' J. Acoust. Soc. Am. **80**, 837–845 (1986).
[26] J. S. Bradley, ''Speech intelligibility studies in classrooms,'' J. Acoust. Soc. Am. **80**, 846–854 (1986).