NRC·CNRC

# NRC Publications Archive
# Archives des publications du CNRC

## Measures for assessing architectural speech security

Gover, B. N.; Bradley, J. S.

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

**Publisher's version / Version de l'éditeur:**
https://doi.org/10.4224/20378300
*Research Report (National Research Council Canada. Institute for Research in Construction); no. RR-171, 2004-11-25*

**NRC Publications Archive Record / Notice des Archives des publications du CNRC :**
https://nrc-publications.canada.ca/eng/view/object/?id=ee925e05-6396-4f1c-8770-0abd4c36c131
https://publications-cnrc.canada.ca/fra/voir/objet/?id=ee925e05-6396-4f1c-8770-0abd4c36c131

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

**Questions?** Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

**Vous avez des questions?** Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

NRC·CNRC CNRC·NRC
Institute for Research in Construction
Institut de recherche en construction
http://irc.nrc-cnrc.gc.ca

# Measures for Assessing Architectural Speech Security

IRC-RR-171

Gover, B.N.; Bradley, J.S.

November 25, 2004

National Research Council Canada
Conseil national de recherches Canada
Canada

## Summary

The issue of architectural speech security is concerned with the degree to which rooms in buildings are acoustically isolated. For example, a conversation occurring within a room being intelligible to persons outside the room is indicative of a problem. To rate the degree to which transmitted speech is intelligible—or even audible—an objective measure is required. The measure should be calculable from the received speech signal spectrum and the background noise spectrum, and it should be highly correlated with responses from actual listeners.

This report describes the design and results of listening tests in which intelligibility scores and audibility ratings were related to a variety of objective measures, calculated from the speech and noise spectra. Existing measures such as Articulation Index (AI), Speech Intelligibility Index (SII), A-weighted level difference, and loudness are included, as well as several different weighted signal-to-noise ratios. The weightings place varying degrees of importance on different frequency bands, as it is known the hearing system response varies with frequency.

The tests involved volunteers listening to a number of English sentences, each filtered to simulate passage through a wall, along with interfering background noise. The subjects responded with the words they were able to understand, or if none, by stating whether the cadence of the speech, or the speech itself, was audible. The results enabled derivation of a new intelligibility measure valid for security, and also investigation of three security thresholds. The *threshold of audibility* defines the point where speech sounds are just barely audible to a listener; the *threshold of cadence* defines the point where the rhythm or cadence of the speech is just recognizable; and the *threshold of intelligibility* is the point where some words are just intelligible.

Pilot tests were performed to assess the importance of many variables. These variables included: talker gender, talker voice characteristics, speech material, speech level, absolute overall level, wall transmission loss characteristics, ambient noise level, ambient noise spectrum, and listener hearing sensitivity. The final tests were designed to be "worst-case", from a security point of view. The participating subjects were excellent listeners, and the speech material was clearly spoken. This "worst-case" approach allows conclusions to be drawn which apply generally: people with less sensitive hearing will not be able to understand more speech than those with good hearing; people with less-clear speaking voices will not be overheard more than clear talkers. The converse situation, testing "median cases", does not allow conclusions to be drawn about louder talkers or better listeners.

It was found that a weighted signal-to-noise measure using the frequency weightings employed in the calculation of SII is an excellent predictor of speech intelligibility scores for security situations. It was also found that, for cases of zero intelligibility, the difference of A-weighted levels is an excellent predictor of audibility of the cadence of speech, or the speech itself. A uniformly-weighted, clipped signal-to-noise measure is a good indicator for all of intelligibility, cadence, and audibility.

## Contents


**Summary** ..... 2
**Contents** ..... 3
**1 Introduction** ..... 4
**2 Background** ..... 5
**3 Test Development** ..... 5
**4 Listening Test Descriptions** ..... 7
*4.1 Test Space and Hardware* ..... 7
*4.2 Speech Material* ..... 10
*4.3 Walls* ..... 11
*4.4 Background Noise* ..... 12
*4.5 Test 1: Index* ..... 13
*4.6 Test 2: Thresholds* ..... 14
*4.7 Subjects* ..... 15
**5 Prospective Numerical Indices** ..... 17
*5.1 Weighted Level Difference* ..... 17
*5.2 Loudness* ..... 17
*5.3 Weighted (Linear) Signal-to-Noise Ratio* ..... 18
*5.4 Weighted (Logarithmic) Signal-to-Noise Level Difference* ..... 19
**6 Results: Intelligibility Scores** ..... 20
**7 Results: Thresholds** ..... 24
**8 Discussion** ..... 32
*8.1 Intelligibility* ..... 32
*8.2 Cadence and Audibility* ..... 35
*8.3 Subject Group* ..... 37
**9 Conclusions** ..... 39
**Acknowledgements** ..... 40
**Appendix A: Pilot Study Summaries** ..... 41
*Subject Response and Scoring* ..... 41
*Multiple Sentences* ..... 41
*Absolute Level* ..... 41
*Different Talkers* ..... 42
*Noise On/Off* ..... 42
**Appendix B: Intelligibility Scoring** ..... 43
**Appendix C: Signal-to-Noise Clipping Level** ..... 46
**References** ..... 48

# 1 Introduction

When occupants of buildings are speaking, their voices can pass through architectural partitions (e.g., walls, doors, floors) into neighbouring rooms and hallways. In many situations, it is important that none of the speech can be understood by persons outside the room containing the talker. In some instances, it may be of further interest to ensure that none of the speech is even audible, or if so, that the rhythm or cadence not be identifiable. These are issues of what can be termed "Architectural Speech Security". Here, "architectural" refers to the security provided by the building structure and associated building systems such as ventilation systems.

Practical questions are those such as "How good a wall/floor system is required to guarantee security?" or "How secure is an existing room?" Before these can be answered, what is required is a method of quantifying the level of security. An objective measure that can address the question "How much of the speech can be understood under given conditions?" must be found.

Defining speech security in terms of the fraction of speech that can be understood casts the problem as one of speech *intelligibility*. This isn't a new idea; investigations into speech privacy for open-plan office situations have used the same approach [1]. The distinction being made is that "privacy" refers to situations free from distraction, where overheard speech can possibly be understood if concentrated upon; "security" refers to situations where overheard speech can *not* be understood, even if concentrated upon. That is, for conditions of high security, the overheard speech is not intelligible.

The existing measures Articulation Index (AI) [2] and its more recent replacement Speech Intelligibility Index (SII) [3] are precisely the sort of measure sought. AI and SII are derived from the speech and noise spectra at the position of a listener. An important characteristic of these measures is that they are based on the *ratio* of signal-to-noise at each frequency, rather than merely on overall speech and noise levels. (Measures of this latter type are commonly used to characterize sound transmission through partitions, but are not good measures of speech security [4].) It is known that different frequencies are of different importance to the understanding of speech, so the speech-to-noise ratio is weighted so as to reflect the importance of the contribution from each band. The contribution to the intelligibility from each frequency band is combined to yield a single number quantity. This quantity (AI or SII) does relate well to the results of many intelligibility listening tests.

However, AI and SII are ill-suited to the problem of security. They do not necessarily represent zero intelligibility at their minimum values of zero. Furthermore, for conditions of zero intelligibility, where the rhythm of speech (or even the speech itself) may or may not be audible, AI and SII provide no information.

Intelligibility tests are used for relating physical characteristics of the sound field (objective measures) to subjective responses of participants (words intelligible to listeners). The usual approach is to present some controlled speech-and-noise combination to a group of listeners and to determine the words that can be understood

(i.e., compute an intelligibility score). A relationship between the intelligibility scores and some objective measure is derived. A useful measure will be well-correlated with the scores, and can then be used to predict intelligibility solely from the sound field measurements. In the present work, tests of this type are used to derive measures that can be used to predict levels of security.

## 2 Background

In the literature, the most significant study of architectural speech "privacy" in buildings was published by Cavanaugh et al. in 1962 [5]. They presented a report on occupants' impressions of privacy in buildings, and how they related to AI. The main focus of the work was on privacy (i.e., freedom from distraction), however there was a subcomponent of the study assessing so-called "confidential privacy", which subjects were instructed corresponded to a situation where there was an "assurance of not being overheard." One of their main results states "the most critical 10% of the subjects began to feel a lack of [confidential] privacy when the articulation index reached 0.05." An important point to bear in mind is that the assessments were based on how private the subjects *felt* a situation was. The actual fraction of the speech they *could* understand was not measured. Nevertheless, this paper certainly lays the basis for relating privacy in offices to an objective measure.

In 1965, Young [6] published a revised computational procedure based on the data in Cavanaugh et al. He proposed a measure derivable from measured A-weighted levels of speech and noise, and single number indicators of transmission loss. This is a much more practical approach, easy to calculate, but at best no more accurate than the original Cavanaugh et al. method.

There have been few other studies published related to architectural speech security [7], though there has recently been interest at international conferences [8,9,10]. There is also ongoing interest among practitioners, due in part to the advent of recent privacy legislation. However, to date no objective measure of speech security has been presented. The present study addresses this issue directly.

## 3 Test Development

There are many types of intelligibility tests. Basic issues to consider when choosing a test method include: presentation method to subjects (headphones, natural listening); subject response method (pen and paper, numeric keypad, oral responses); test material (words, syllables, sentences, extended discourse); and scoring procedure (numerical scoring, subjective estimation or rating). To a large extent the response method and scoring procedure depend on the test material.

The intention of the present tests was to simulate listening outside of a room, so it was decided that subjects would sit and listen to sound fields presented over loudspeakers. Speech material was filtered to make it sound like it had passed through a wall, and then was played to subjects, along with background noise. This simulated the acoustical

conditions outside of a room in which someone was talking. The subjects listened to the sounds, and then were asked to respond to what they heard.

**Speech Material/Scoring** The test material used in the present tests was short sentences. Sentences are more representative of conversation than isolated words, and if constructed carefully, can be of low predictability. The response and scoring procedure had the subjects say out loud the words they thought they heard, and the test operator noted on a score sheet those that were correctly identified. In this way the fraction of words actually understood was scored. In intelligibility tests using rhyming words, the scoring can be done similarly, but subjects are able guess correctly as often as 20% of the time [11], which means they are particularly unsuitable in assessing security situations, where scores as low as 0% are expected. Furthermore, sentences can be more accurately scored than running speech, or so-called "connected discourse", whereby the procedure usually involves allowing the subject to estimate the fraction of the words they are capable of understanding [12].

**Talker** The same sentence spoken by different talkers—or even by the same talker with different effort or speaking style—will have a different spectral energy distribution. For instance, it is known that on average, male voices have a larger proportion of low-frequency energy than female voices, and that for a given speaking effort, male voices have higher A-weighted levels [5]. This could be interpreted as leading to a varying degree of security for a given "architecture" (wall, room, etc.), depending on the talker. A pilot test (see Appendix A) was conducted repeating an identical test for nine different recordings of the test sentences. The voice that was most easily heard by the subjects was a male, speaking clearly, and also happened to be the highest-quality recording (from a digital audio point of view). This is the voice used in the main tests, and is described in Section 4.2.

**Walls** As stated above, the sentences were each filtered to simulate transmission through a wall. A number of wall transmission loss (TL) curves were selected from past measurements of actual wall samples in the testing facility in Building M-27 at NRC. Various wall constructions were considered, including doors, partitions plus ceiling plenums, and wood and steel stud wall assemblies made with gypsum board, and absorbing materials in the cavities. Each modifies the speech differently, according to the frequency dependence of the TL. Four different TL curve shapes were selected from the set of candidate walls, as discussed in Section 4.3.

**Noise** Background noise greatly influences intelligibility of speech. The nature of the noise (random, impact, periodic, talking, etc.) is very important. For the present tests, "ventilation-type" random noise was used, described in Section 4.4. The nature and spectra of these noises are representative of background noises in offices.

It is often casually observed that persistent background noise is not really noticeable until it shuts off. To investigate whether this acclimatization had an effect on the intelligibility tests, a pilot test was conducted, discussed in Appendix A. There was no significant

effect of leaving the noise on in between test sentences, as compared to switching it on and off before and after each sentence.

**Absolute Level** One approach for setting up test conditions is to say that for a given wall and a given speaking voice level, the background noise of a particular spectral shape is to be varied in level, thereby producing a range of intelligibilities. When the noise is (relatively) quiet, the speech will be most intelligible; when it is loud, the speech will be least intelligible. The range through which the noise level must be adjusted can be great, and in some cases may be unrealistic. For instance, for a particularly "bad" wall, the noise may have to be raised to 70 dBA to sufficiently mask the speech, which is unreasonable. To get around this problem, what can be done is to lower both the speech and noise by the same amount, preserving the same *ratio*, but changing the overall absolute level. The question is whether this will change the intelligibility of the test sentence. To investigate this problem, a pilot test was conducted including cases with a fixed signal-to-noise ratio, but varying absolute level (see Appendix A). There was no statistically significant effect observed. Effort was maintained throughout the testing to maintain reasonable source speaking levels and background noise levels.

# 4 Listening Test Descriptions

Individual subjects were seated in a sound-isolated room and exposed to sound fields simulating various "speech outside of a room" conditions. One "condition" corresponded to one combination of: talker voice level, wall transmission loss, and background noise spectrum and level. All tests were approved by NRC's Ethics Review Board.

## *4.1 Test Space and Hardware*

All tests were conducted in the Room Acoustics Test Space in Building M-59 at the National Research Council. This is a room measuring 9.2 m long by 4.7 m wide by 3.6 m high, constructed from concrete. The room is not connected to the building, and is resting on springs for vibration isolation. Sounds existing outside the room, therefore, are largely isolated from penetrating within. For the present study, the interior walls of the room were lined with 10 cm-thick absorbing foam, covered by curtains. There was a conventional T-bar ceiling with 25 mm-thick glass fibre ceiling tiles installed, and the floor was covered with carpet. This interior treatment yielded a quite "dead" space, enabling complete control of the sounds within the room. The background noise level in the room was about 13.7 dBA (measured with the sound simulation system turned off). Figure 1 shows a measurement of the background noise spectrum.

The test speech was played over loudspeakers positioned at the front of the room. The background noise was played over another set of loudspeakers positioned above the ceiling, directly above the subject, which is where ventilation noise would typically originate. Figure 2 shows a diagram of the setup. Having the noise and speech originate from different spatial locations is important for a realistic test. Listeners are better able to ignore the noise and focus on the speech than if both originated from the same direction [13].

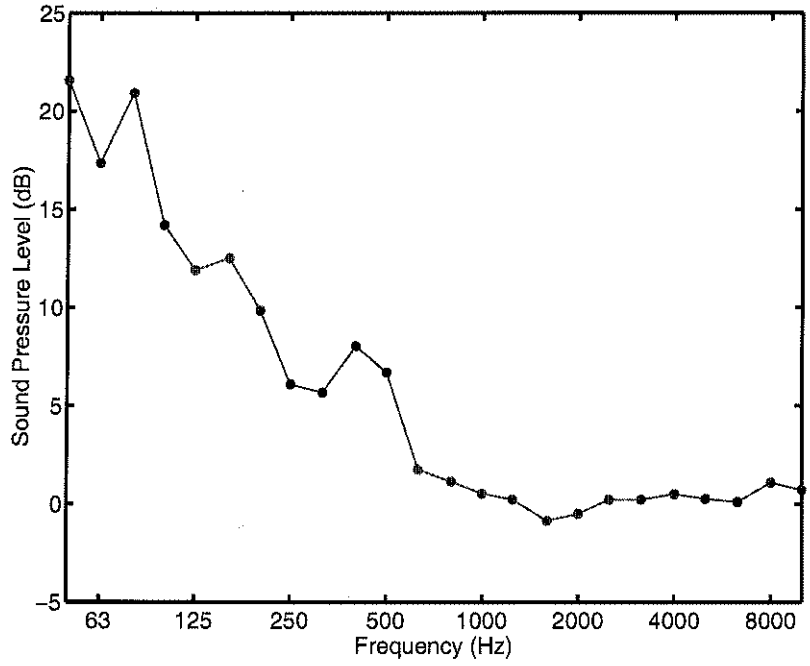


**Figure 1 Measured background levels in the Room Acoustics Test Space. The A-weighted level of this measurement is 13.7 dBA.**

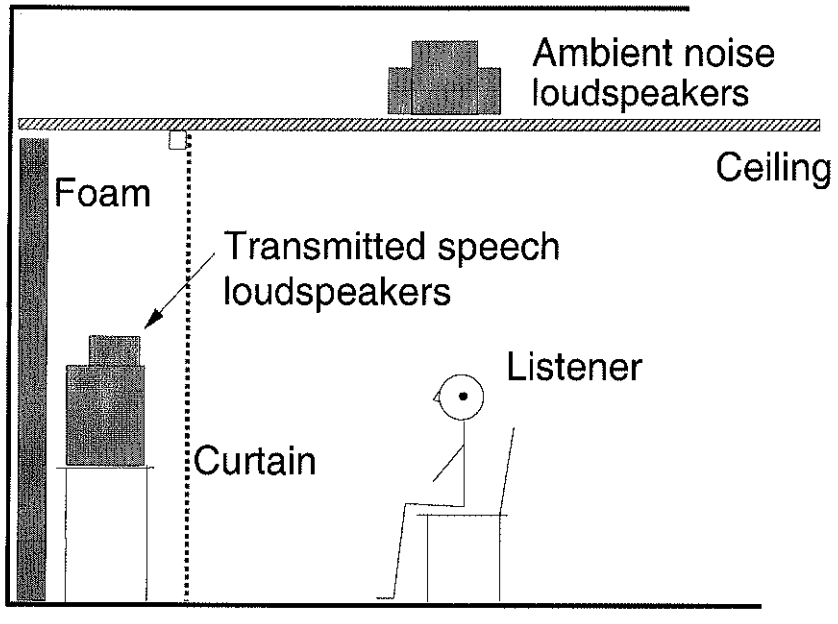


**Figure 2 Schematic of cross-section through Room Acoustics Test Space showing location of the listener and the loudspeakers used to generate the test sound fields.**

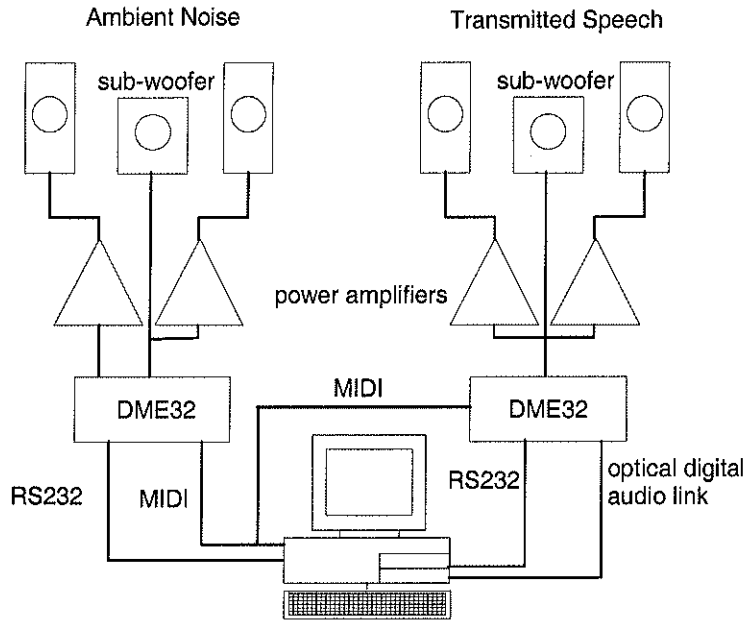


**Figure 3 Block diagram of the computer controlled electroacoustic system used to create the test sounds.**

A block diagram of the electroacoustic system used to produce the test sounds is shown in Figure 3. The two blocks labeled "DME32" are Yamaha Digital Mixing Engines, which are highly flexible signal processing boxes, able to perform the functions of many interconnected devices such as equalizers, filters, oscillators, etc. The output signals of the DME32 units are fed through the power amplifiers into high-quality loudspeaker systems (Paradigm Compact Monitors, Paradigm PW sub-woofers). One component in each DME32 was initially configured under computer control (via the RS232 interface) to equalize the playback path through the power amplifiers and loudspeakers to be flat at position of the listener's head (± 1 dB from 60 to 12000 Hz).

The background noises for the test sound fields were generated by the internal noise generator of one of the DME32 units. This same unit shaped the spectrum and adjusted the level as desired, responding to control commands sent by the computer over the MIDI interface. One channel of the noise output was delayed by 300 ms relative to the other so as to avoid any unnatural perceptual effects caused by movement of the listener's head.

The speech sounds were generated from playback of anechoically-recorded source material stored on the computer in 16-bit, 44.1 kHz wave-file format. The output of the sound card ran into the second DME32, which performed the necessary equalization and level adjustment, again with commands to switch settings received over the MIDI interface. Equalizer components in this DME32 simulated the various wall transmission loss curves.

## 4.2 Speech Material

The set of test sentences used is known as the "Harvard Sentences", and is made up of 72 phonetically-balanced lists of 10 sentences each (720 different sentences) [14]. These sentences are not "nonsense"; they are all correct English sentences that make sense. They are, however, highly unpredictable, which means that even understanding the first few words does not greatly increase the chances of guessing the rest of the sentence. They are therefore highly-suited to the present application.

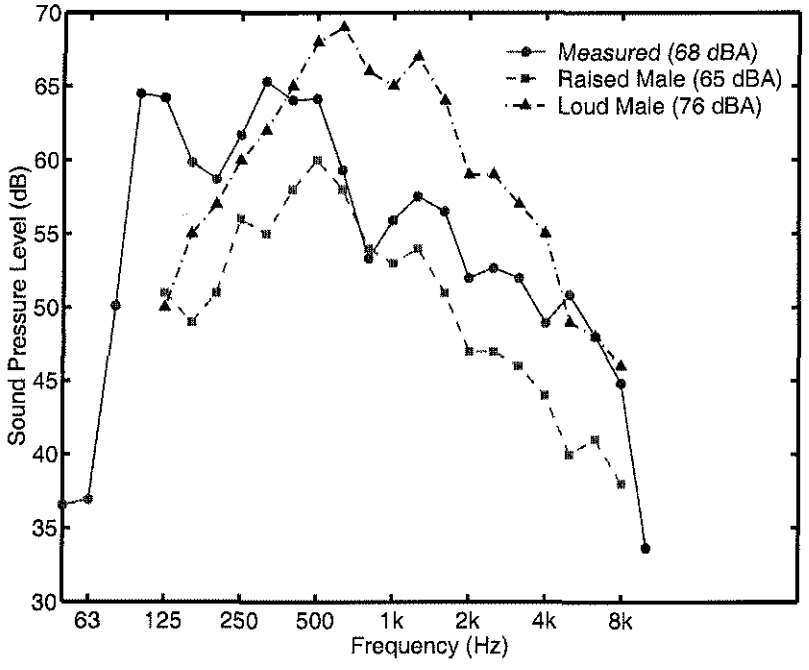


**Figure 4 Measured spectrum of talker used in this work. Also shown are "typical" spectra for a male talker speaking with "Raised" and "Loud" effort [15].**

Nine recorded versions of the test sentences were acquired, spoken with different efforts by different talkers of both genders. The choice of recording used for the tests was the best quality recording available (16-bit, 44.1 kHz "CD-quality" digital), and was of a male talker speaking clearly. This speaking voice was most intelligible to subjects in pilot testing. That is, from a security point of view, this recording was the "worst-case" talker. An average (of measurements of 4 different sentences) of the magnitude spectrum of the speech is shown in Figure 4, along with "typical" speech spectra for a male talker speaking with two different speaking efforts, taken from Ref. [15]. At this playback level, it can be seen that the speech corresponds to an effort somewhere between "Raised" and "Loud". All speech spectral levels were measured over a 60 second period, looping the sentences continuously.

## 4.3 Walls

Wall transmission loss (TL) curves were selected from past measurements of actual wall samples in the wall testing facility in Building M-27 at NRC. The TL curves for the four walls used in the test are shown in Figure 5, their descriptions are shown in Table 1. These four curves were selected since they are of different shapes, and representative of wall constructions typical of office environments.

The speech transmitted through any of these walls will be attenuated and spectrally-distorted (i.e., "muffled"). Figure 6 shows typical speech spectra measured in the Room Acoustics Test Space after filtering by each transmission loss curve.

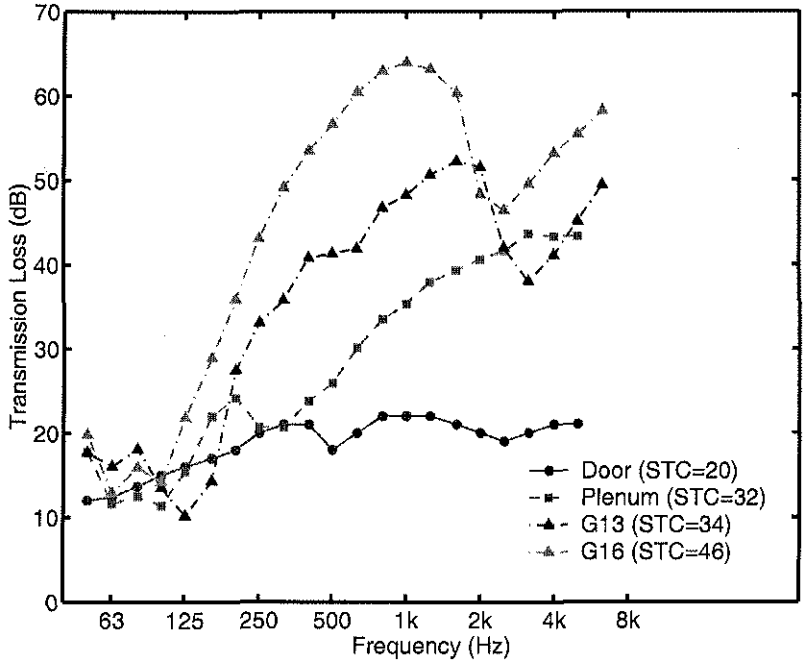


**Figure 5 Wall transmission loss curves, obtained through measurements in the M-27 wall test facility.**

| *Wall Descriptor* | *Wall Description* | *STC Rating* |
|---|---|---|
| Door | Solid core door, no seals. | 20 |
| Plenum | 5/8" mineral fibre ceiling. | 32 |
| G13 | 89 mm wood stud wall with 13 mm gypsum board on both sides; cavity filled with glass fibre batts. | 34 |
| G16 | 90 mm steel stud wall with 16 mm gypsum board on both sides; cavity filled with glass fibre batts. | 46 |

**Table 1 Wall descriptions and sound transmission class (STC) ratings.**

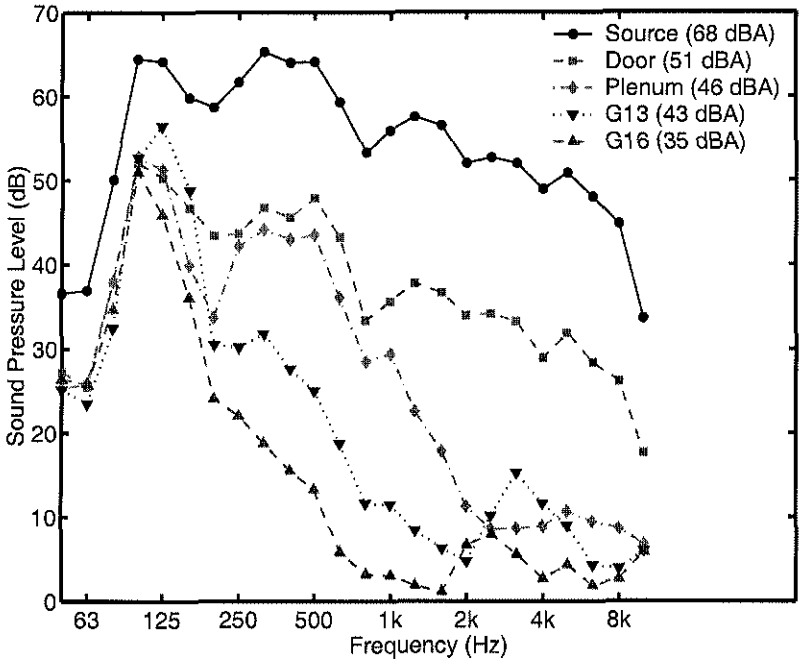


**Figure 6 "Transmitted" speech spectra, measured in the test room after filtering the source speech (top curve) through each simulated wall.**

## *4.4 Background Noise*

All background noises used in the testing were spectrally-shaped random noise. As discussed above, one of the Yamaha DME32s was used to generate the noise internally, and equalizer components were used to shape the spectrum. Five different spectra were used, shown in Figure 7.

The base case noise is the "Neutral" spectrum, so-named under the RC naming convention [16]. It has a –5-dB/octave roll off. The other spectra are derived from this by boosting the "low" (50–200 Hz), "mid" (250–1600 Hz), or "high" (2000–10000 Hz) frequency sections. These other noise types are used to systematically vary the spectral qualities of the interfering noise, representing cases which are more or less "rumbly" or "hissy".

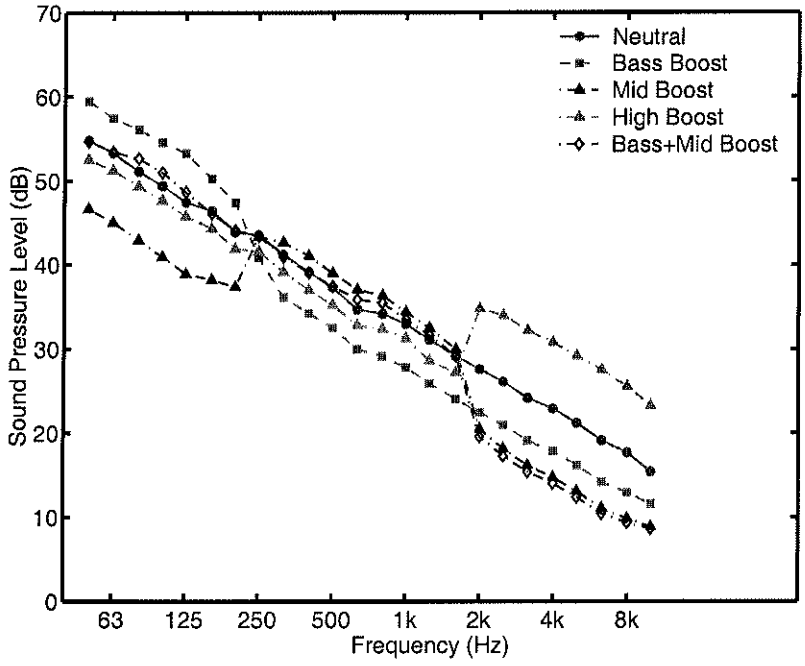


**Figure 7 Measured noise spectra, all corresponding to 45 dBA.**

## 4.5 Test 1: Index

This test consisted of 340 sentences, processed to simulate wide ranges of the conditions thought to be representative of actual offices and meeting rooms. Each of the four wall types and five background noise types was included. For each wall/noise combination, three signal-to-noise levels were designed, corresponding to varying "difficulty": easy, moderately difficult, and difficult. In the "easy" case it was judged that most listeners should be able to identify all words in a sentence. The "difficult" case was intended to be just above the threshold of intelligibility—some listeners could identify some (but not necessarily all) of the words. It is not possible to design in advance the conditions that define the threshold we are seeking to find. The difficult cases still had to correspond to "some intelligibility" in order to find the threshold point—from a scoring point of view, zero intelligibility just below the threshold is the same as zero intelligibility 100 dB below the threshold. In addition to these 4 x 5 x 3 = 60 cases, 8 additional cases were constructed corresponding to 2 additional difficulties ("very difficult" and "very very difficult") for 1 noise type, for all 4 walls. These cases were constructed by reducing the speech level 3 dB and 9 dB below that used for the "difficult" case. In total, this yielded a total of 68 physical conditions. Five different sentences were included for each condition, resulting in a test 340 sentences long.

The test was conducted in the following manner: the subject sat in the room with no test sounds playing. The noise was turned on, and a second or so later (the timing was variable), the speech began. After the end of the sentence, the noise was switched off. At this point, the subject said out loud the words they thought they heard. The subjects were encouraged to guess. The operator was outside the test facility scoring the

responses, and would ask the subject to repeat an answer deemed ambiguous or incomprehensible. When the operator was satisfied, the computer was cued to play the next test sentence.

Subjects completed the test over three different testing sessions, usually on different days. Each session consisted of two runs of about 57 sentences, separated by a brief break to avoid fatigue. To assuage "learning onset", the subjects listened to several practice sentences each time. In total, one run of 57 sentences took about 20 minutes; the session for the day taking just under an hour.

Table 2 shows a portion of the score sheet used to tally responses. The response was recorded by highlighting the words correctly repeated by the subject. The score was tallied both by counting all words in the sentence (the "Full" score), and also counting only so-called "key" words in the sentences (the nouns, verbs, etc.) [17]. This was done because it can be argued that hearing only the non-key (a.k.a. "connective") words probably means that the information being conveyed by the sentence is not being understood by the subject. For instance, for sentence 01 in Table 2, if the subject repeats "*Something and something as the something something.*", then the Full score is 3/8, but the Keyword score is 0/5. It is up to security designers to decide whether such a situation is acceptable or not. (Another strategy is to use a weighted score whereby keywords are worth 1 point and connective words are worth ½ point. This can be calculated by simply averaging the Full and Keyword scores.) The analyses in this report contemplate only the "Full" scores, assuming that any single word is important. For completeness, however, some analyses for the other scoring strategies are presented in Appendix B.

| | Sentence | Not Audible | Keyword | Full |
|---|---|---|---|---|
| 01 | *Yell* and *clap* as the *curtain slides back.* | | / 5 | / 8 |
| 02 | *They* are *men* who *walk* the *middle* of the *road.* | | / 5 | / 10 |
| 03 | *Both brothers wear* the *same size.* | | / 5 | / 6 |
| 04 | In *some form* or *other* we *need fun.* | | / 5 | / 8 |
| 05 | The *prince ordered* his *head chopped off.* | | / 5 | / 7 |

**Table 2 Sample of score sheet for the 340-sentence intelligibility test.**

## 4.6 Test 2: Thresholds

A second test was conducted as a follow-up to the previous one, consisting of an additional 160 test sentences. These resulted from the 32 combinations of 2 wall types (G13 and G16) x 2 noise types (Neutral and Bass Boost) x 8 signal/noise levels, with 5 different sentences for each combination. These cases were distributed over the range of difficulty from all subjects able to understand at least one word, through to all subjects unable to detect the presence of speech at all. The "best" listeners from the previous test were used as subjects. ("Best" means these subjects correctly identified the most words, that is, they are the "worst-case" listeners from a security point of view.)

The test was conducted in a manner exactly like the previous one, except in the manner in which the subjects responded. They still spoke out loud, but were asked to respond to the following questions:

- Did you hear any speech sounds?
- If yes, did you hear the rhythm or cadence of the speech?
- If yes, did you understand any of the words? (Tell the experimenter the words you were able to understand.)

A sample of the score sheet for this experiment is shown in Table 3. For each sentence, the response was either: Not Audible (heard no speech sounds), or No Cadence (heard speech sounds, but not the rhythm), or Heard Cadence (heard speech cadence, but identified no words correctly), or some number (>0) of words correctly identified.

| | Sentence | Not Audible | No Cadence | Heard Cadence | KeyWord Score | Full Score |
|---|---|---|---|---|---|---|
| 01 | A *king ruled* the *state* in the *early days.* | | | | / 5 | / 9 |
| 02 | The *ship* was *torn apart* on the *sharp reef.* | | | | / 5 | / 9 |
| 03 | *Sickness kept* him *home* the *third week.* | | | | / 5 | / 7 |
| 04 | The *wide road shimmered* in the *hot sun.* | | | | / 5 | / 8 |
| 05 | The *lazy cow lay* in the *cool grass.* | | | | / 5 | / 8 |

**Table 3 Sample of score sheet for the 160-sentence threshold test.**

## 4.7 Subjects

Subjects participating in the tests were respondents to a broadcast email at NRC, plus some additional participants from PWGSC. All were fluent English speakers, and none was compensated for participating. All respondents were given a standard audiometric hearing threshold test and a short trial of the intelligibility test. The 36 subjects correctly identifying 62% or more of the words in the trial participated in the main test (Test 1). Of these, 16 were male and 20 were female. The mean intelligibility score (across all 340 sentences) for each of these 36 subjects was computed and is shown in Figure 8. This was used to split the group into the "better" subjects (the 19 scoring greater than 65%, of whom 7 were male and 12 were female) and the "worse" subjects (the 17 scoring less than 65%).

These "better" 19 subjects additionally participated in Test 2. The hearing level (HL) curves for the subjects are shown in Figure 9. The average HL for frequencies 0.5, 1, 2, 3, and 6 kHz for both ears can be used to categorize the subjects. This quantity is denoted F-HL (frequency-averaged hearing level), and was equal to 0.26 dB with a standard deviation of 3.42 dB for the 19 "better" subjects, and equal to 5.62 dB with a standard deviation of 4.74 dB for the entire group of respondents. The median otologically normal female listener has a F-HL of about 0 dB at age 20, about 1.2 dB at age 30, about 3.4 dB at age 40, and about 6.8 dB at age 50 [18].

It should be noted that factors other than measured HL contribute to a subject's proficiency at identifying the words. Attention span, ability and desire to concentrate and focus, and native language play roles. It is not understood to what degree each factor influences performance.

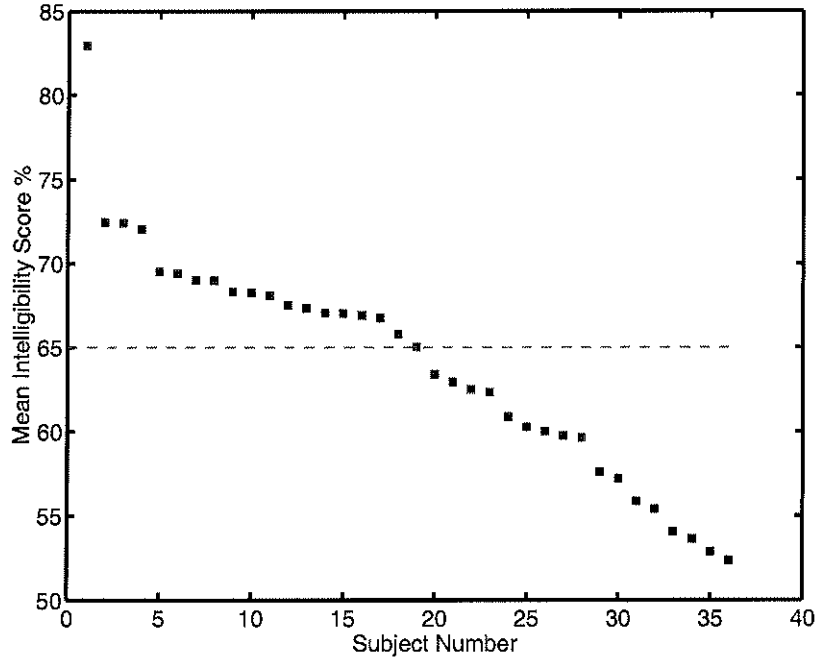


**Figure 8 Mean subject intelligibility scores for the 340-sentence test. The dashed line at 65% is used to divide the group in two.**

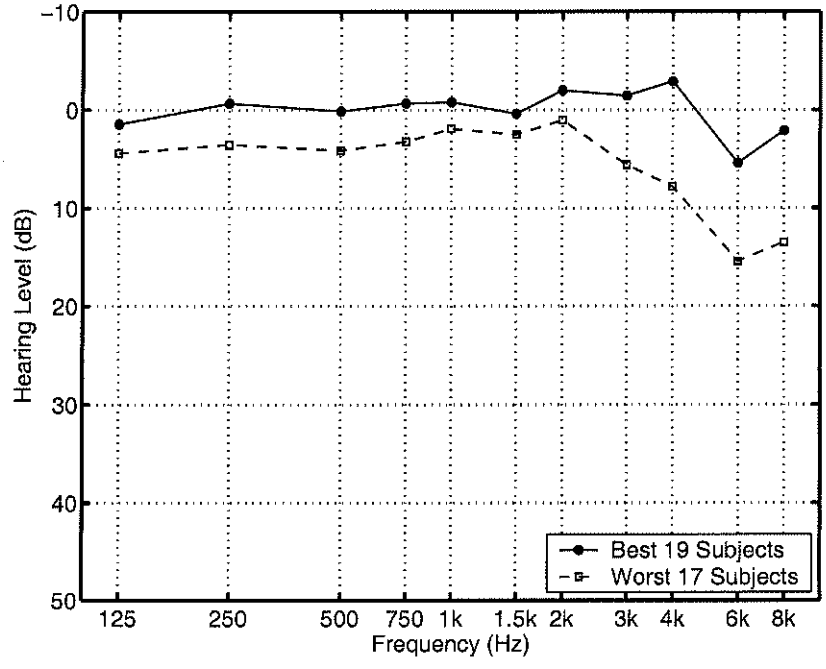


**Figure 9 HL curves for the "better" and "worse" groups of subjects.**

# 5 Prospective Numerical Indices

From measurements of the third-octave spectra of the speech and noise, various numerical indicators can be calculated. The issue is to find a measure well-correlated with the results of the listening tests described in Section 4. In general terms, there are two categories of measures of interest: 1.) summing signal and noise spectra separately, then computing the difference in the result, or 2.) computing the ratio of the signal and noise spectra, then summing the result. These two categories lead to many distinct indicators by differing choice of frequency weightings in the summations (applying different importance to different frequency bands).

## 5.1 Weighted Level Difference

The speech and noise levels can be computed from their spectra separately, and then the difference in these levels computed. The expression for the weighted level difference $L_w$ is

$$L_w = 10\log_{10}\left(\sum_b w_b S_b\right) - 10\log_{10}\left(\sum_b w_b N_b\right), \tag{1}$$

where $S_b$ is the speech intensity (not in dB) in frequency band $b$, $N_b$ is the noise intensity (not in dB) in band $b$, and $w_b$ are the frequency weightings. The terms on the right-hand side are the weighted signal and noise levels, respectively.

The main weighting scheme considered as a potential indicator is the familiar A-weighting, the result being the *A-weighted level difference* [19].

## 5.2 Loudness

*Loudness* is a quantity that can be calculated from the spectrum of a signal, and has been shown to be related to the perceived loudness of sounds. The calculation of loudness is explained in Refs. [20] and [21]. The unit of loudness is the *sone*, which is a linear measure: a signal with double the loudness of a second signal will sound twice as loud. Measuring the speech loudness $\Lambda_S$ and the noise loudness $\Lambda_N$ separately, the loudness ratio $r_\Lambda$ given by

$$r_\Lambda = \frac{\Lambda_S}{\Lambda_N}, \tag{2}$$

can be computed. For instance, if the speech is half as loud as the noise, the value of $r_\Lambda$ will be 0.5. Loudness (in sones) was used, rather than *loudness level* (in dB), since the computational procedure for loudness level involves first computing loudness anyway. The relationship between the two is non-linear below 40 sones (as all the speech signals in the tests were), and the conversion is non-trivial. The linear measure loudness is well-defined in any case.

## 5.3 Weighted (Linear) Signal-to-Noise Ratio

A different approach is to compute the signal-to-noise ratio in each frequency band first, and then perform a (weighted) sum across all frequencies. This is similar to some of the steps involved in computing AI and SII (which are outlined in Refs. [2] and [3]). The expression for the weighted signal-to-noise ratio $R_w$ is

$$R_w = 10\log_{10}\sum_b w_b\left(\frac{S_b}{N_b}\right), \tag{3}$$

where $S_b$ and $N_b$ are as above. The weights $w_b$ in this case reflect the importance of the *ratio* of speech-to-noise in each frequency band $b$.

Weighting strategies considered are given in Table 4, they are plotted in Figure 10. In the figure, the weights are plotted in dB, and are all normalized to have a peak of 0 dB. (The correlation weightings shown were derived from analysis of the intelligibility scores, they will be different for the analyses of the thresholds.)

| *Weighting* | *Description* |
|---|---|
| Uniform | Equal importance for all frequencies. |
| A | A-weighting. |
| AI | Weights used in calculation of AI. |
| SII | Weights used in calculation of SII. |
| Low-Frequency (LF) | Increased importance for low-frequency bands relative to higher-frequency bands. Implemented as a 1.5 dB per octave roll-off. |
| Correlation | Weights proportional to the square of the correlation coefficient ($R^2$) of individual frequency band signal-to-noise ratio with test scores. |

**Table 4 Descriptions of signal-to-noise weighting strategies.**

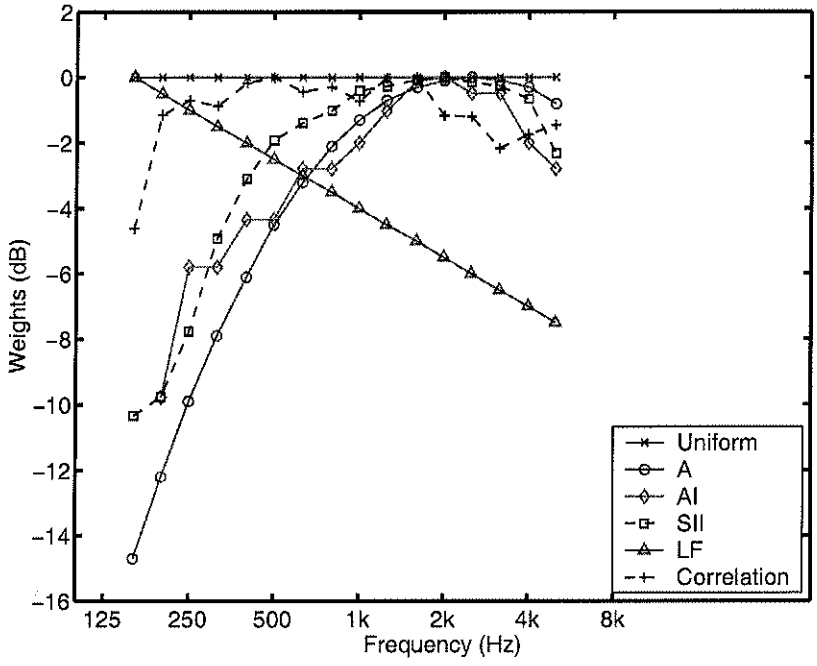


**Figure 10 Frequency weighting curves in decibels, corresponding to the weighting strategies given in Table 4.**

## 5.4 Weighted (Logarithmic) Signal-to-Noise Level Difference

The above weighted sum of signal-to-noise ratios $R_w$ is a little unconventional. In calculation of AI and SII, the weighted summation is of signal-to-noise *level differences* (in dB). To maintain consistency with convention, therefore, we can define an index that is a simple weighted sum of band signal-to-noise level differences.

$$X_w = \sum_b w_b \left(10\log_{10} S_b - 10\log_{10} N_b\right), \tag{4}$$

where $S_b$ and $N_b$ are as above. $10\log_{10} S_b$ and $10\log_{10} N_b$ are the speech and noise levels (in dB), respectively, in band $b$. One problem that arises with this summation is in cases of very little signal, the signal-to-noise level difference blows up (approaches negative infinity). These large, negative numbers can unrealistically skew the summation. This is not a problem in the linear ratio summation, since the ratio merely approaches zero in such a case. One solution is to clip the band signal-to-noise level difference at some minimum value. All values of $\left(10\log_{10} S_b - 10\log_{10} N_b\right)$ less than some minimum value $L$ are set equal to $L$. This is in fact part of the AI calculation procedure (with $L = -12$ dB) and of the SII calculation procedure (with $L = -15$ dB).

Weighting strategies considered for this measure were: uniform, AI, and SII (all as above in Sec. 5.3). Level difference clipping levels are discussed in Appendix C. Levels of $L = -22$ dB and $L = -32$ dB are employed for analysis.

# 6 Results: Intelligibility Scores

Figure 11 shows the results of the intelligibility tests, plotted versus the pre-existing indices (A-weighted level difference, loudness ratio, AI, and SII) and the new weighted (linear) signal-to-noise ratio indices (Sec. 5.3). Figure 12 shows the same results, plotted versus the pre-existing indices and the new dB-summed weighted signal-to-noise level differences (Sec. 5.4). Each panel of both figures shows the 19 x (340 + 160) = 9500 individual intelligibility scores for the "better" 19 subjects for each sentence, as a fraction of the words correctly identified plotted versus various objective indicators computed from the noise and speech spectra (the dots). Overlaid on the plots are the least-squares best-fit Boltzmann curve [22] to the mean of the data (the dashed curve), and fits to the $5^{th}$, $50^{th}$, and $95^{th}$ percentile curves (solid curves). The $N^{th}$ percentile curve is defined so that (100-N)% of the subjects have intelligibility scores that exceed the values on the line; that is, (100-N)% of the data points lie above the line, N% lie below. The square of the correlation coefficient $R^2$ (also called the "coefficient of determination") of the relationship between the scores and the index is given.

It is worth noting that this data is not easily handled by typical regression methods [23]. Being fractions, it is not possible for the scores to be less than 0 or greater than 1. This means a fitting curve must be non-linear; it must have asymptotes at 0 and 1. Furthermore, the data is non-normally distributed about this best-fit curve. For example, at high values of intelligibility (near 1), data points can fall farther "below" the curve than "above" it. For these reasons, the data was analyzed in narrow "windows" (on the horizontal axis), and the cumulative distribution and percentiles computed within each. The smooth curves shown on the figure are the Boltzmann function fits to the percentiles in each window. An analogous procedure was followed, and explained in more detail, in Ref. [24].

Nonetheless, the squared correlation coefficients can be used to evaluate the goodness of the indicators as predictors of intelligibility. Values closer to unity indicate better relationships, closer to zero indicate worse ones. The $R^2$ for the A-weighted level difference (0.464) and the loudness ratio (0.565) are the worst. AI and SII have $R^2$ values of 0.726 and 0.745, respectively, which are among the highest. Do notice, however, that AI and SII themselves do have the problem that at their minimum value of zero, subjects are still able to correctly identify words from the test sentences. They are therefore confirmed to be unsuitable as a security measure.

All of the other indices in Figure 11 are weighted signal-to-noise (linear) ratios, the only difference being in the weighting. The largest $R^2$ values are for the AI (0.747) and SII (0.744) weightings, and the A-weighting (0.739). Weighting according to the band-by-band squared correlation coefficient yields a slightly lower value (0.711). Uniform weighting (0.670) and LF-weighting (0.599) result in worse correlations.

The main qualitative difference among the weighting strategies that perform well (AI, SII, A) and the others is in the lower frequencies. The uniform and LF weightings place greater relative importance on frequency bands below about 800 Hz than do the others. Evidently these frequencies are not as important for intelligibility.

The other six indices presented in Figure 12 are weighted signal-to-noise level differences: three different weightings (uniform, AI, SII) for each of two different signal-to-noise clipping levels ($L$ = –22 and –32 dB). The highest $R^2$ values are for the SII-weighting, $R^2$ = 0.762 for –22 dB clipping, and $R^2$ = 0.757 for –32 dB clipping. The AI-weighted indices yield $R^2$ = 0.755 and $R^2$ = 0.750 for –22 dB and –32 dB clipping, respectively. The uniformly-weighted indices result in $R^2$ = 0.750 for both –22 and –32 dB clipping. (These clipping levels were selected after having assessed the relationships between the test results and the indices for clipping from –12 dB down to –32 dB, in 2 dB steps. These results are discussed in Appendix C.)

All of the weighted signal-to-noise indices presented in Figure 12 are well-correlated with the intelligibility scores. The differences among the $R^2$ values are statistically significant, but it is not clear that they are all practically significant. In general, the SII-weighted measures correlate slightly better than the others, the highest for –22 dB clipping of the signal-to-noise level difference.

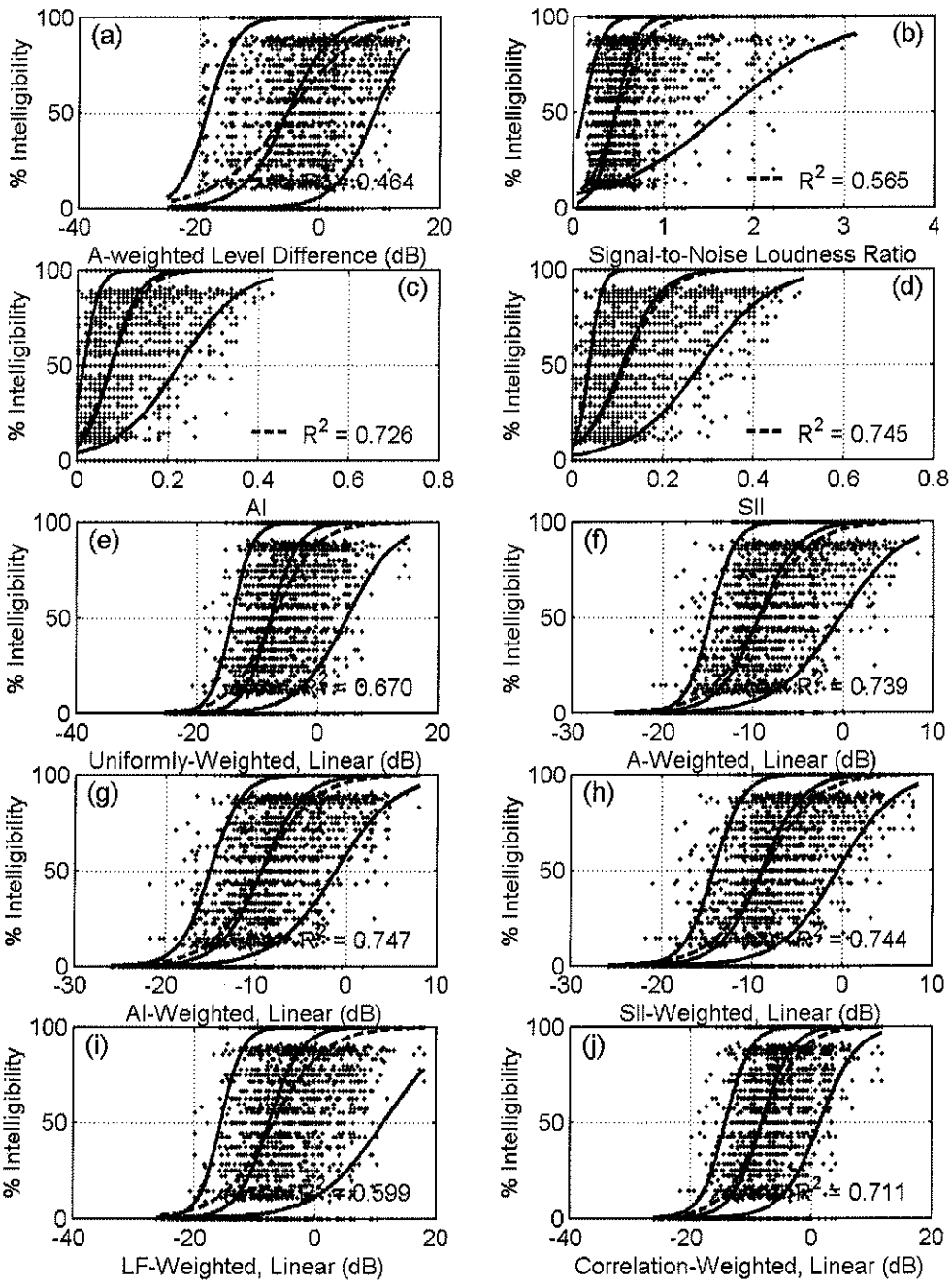


**Figure 11 Linear signal-to-noise ratio-summed measures: Individual intelligibility scores for 19 subjects on 500 sentences each (9500 points total), plotted versus various indices. The dashed line is the least-squares Boltzmann function fit to the mean. The three solid lines are the least-squares Boltzmann function fits to the 5$^{th}$, 50$^{th}$, and 95$^{th}$ percentiles (from bottom to top, respectively).**

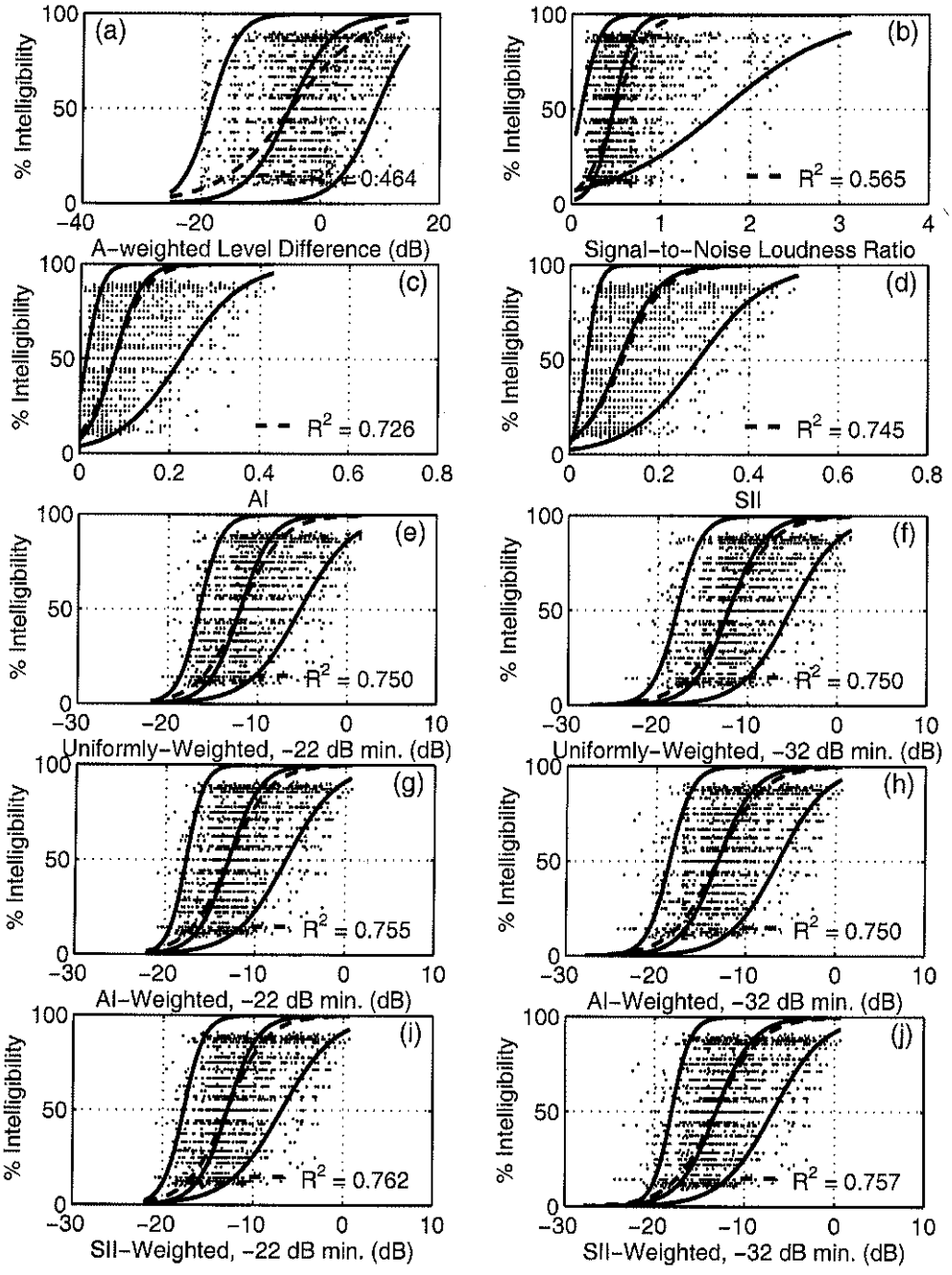


**Figure 12 Decibel level difference-summed measures: Individual intelligibility scores for 19 subjects on 500 sentences each (9500 points total), plotted versus various indices. The dashed line is the least-squares Boltzmann function fit to the mean. The three solid lines are the least-squares Boltzmann function fits to the 5$^{th}$, 50$^{th}$, and 95$^{th}$ percentiles (from bottom to top, respectively).**

# 7 Results: Thresholds

The analysis for the threshold data is similar to that for the intelligibility scores. For each of the 340+160 sentences in both tests, a tally was made of the fraction of the 19 subjects able to correctly identify at least one word. These data are plotted in Figure 13 and in Figure 14. This is the Threshold of Intelligibility analysis.

Notice that as above for the intelligibility word scores, A-weighted level difference and loudness ratio are the poorest indicators ($R^2$ = 0.586 and 0.737, respectively). Also as above, the SII-weighted, AI-weighted, and A-weighted (linear) signal-to-noise ratios had the highest $R^2$ values (0.925, 0.915, and 0.919, respectively). The $R^2$ values for AI (0.889) and for SII (0.904) were slightly lower. Those for the other linear ratio weighting schemes were: 0.864 for uniform weighting, 0.900 for correlation-weighted, and 0.783 for LF-weighted. From Figure 14, the –22 dB clipped signal-to-noise ratios had the highest $R^2$ values (0.919, 0.910, and 0.908 for the SII, AI, and uniform weights, respectively). The $R^2$ for the –32 dB clipped schemes were: 0.905 for SII weighting, 0.896 for AI weighting, and 0.900 for uniform-weighted. (Appendix C contains a discussion for additional clipping levels.)

For the 160 sentences of Test 2, a tally was made of the fraction of the 19 subjects able to: 1.) identify the cadence or rhythm of the speech (including those identifying words), and 2.) able to hear the presence of speech in the background noise (including those identifying cadence or words). These results are shown in Figure 15 and Figure 16, for the analysis of the Threshold of Cadence, and in Figure 17 and Figure 18 for the Threshold of Audibility.

For the threshold of cadence analysis, $R^2$ was 0.918 for the A-weighted level difference and 0.956 for the loudness ratio. The $R^2$ values for AI and SII were 0.672 and 0.770, respectively. For the linear ratio summed indices, notice that the uniformly-weighted and LF-weighted indices are better predictors than the A-weighted, AI-weighted, and SII-weighted signal-to-noise indices, the converse of what was observed for the intelligibility case. The –22 dB clipped measures had $R^2$ values of 0.912, 0.798, and 0.815 for uniform, AI, and SII weighting, respectively. The –32 dB clipped measures had $R^2$ values of 0.858, 0.686, and 0.691 for uniform, AI, and SII weighting, respectively. (Appendix C contains a discussion for additional clipping levels.)

For the threshold of audibility analysis, $R^2$ was 0.835 for the A-weighted level difference, 0.899 for the loudness ratio. The $R^2$ values for AI and SII were 0.389 and 0.566, respectively. The highest correlations among the linear ratio summed indices were for LF-weighted (0.819), uniformly-weighted (0.808), and correlation-weighted (0.777). The $R^2$ values for the A, AI, and SII weightings were 0.659, 0.670, and 0.701, respectively. The –22 dB clipped dB-summed measures had $R^2$ values of 0.816, 0.681, and 0.693 for uniform, AI, and SII weighting, respectively. The –32 dB clipped measures had $R^2$ values of 0.741, 0.581, and 0.583 for uniform, AI, and SII weighting, respectively. (Appendix C contains a discussion for additional clipping levels.)

In general, for both cadence and audibility thresholds the A-weighted level difference and loudness ratio are better indicators than the weighted signal-to-noise ratio type measures, including AI and SII. The thresholds of cadence and of audibility involve the detection of sounds, not necessarily understanding of speech. They are more related to audibility and perceived loudness than to intelligibility. The strong correlations with loudness (the measure) and A-weighted level difference, which are essentially measures of audibility, indicate this.

For these two thresholds (cadence and audibility), the –22 dB clipped measures were better-correlated than the –32 dB clipped, but inspection of the figures indicates that the best-fit curves do not reach zero. This is analogous to the problem with AI and SII for intelligibility scores. Deciding which measure is best requires inspection of more than just the $R^2$ of the relationship.

For both thresholds, the uniformly-weighted measures are better-correlated than the AI and SII weighted ones. This possibly indicates the relative importance of the lower frequencies (below 800 Hz) for audibility, which are reduced by the intelligibility-derived weighting schemes, but are relatively important for transmitted speech sounds (see Figure 6).

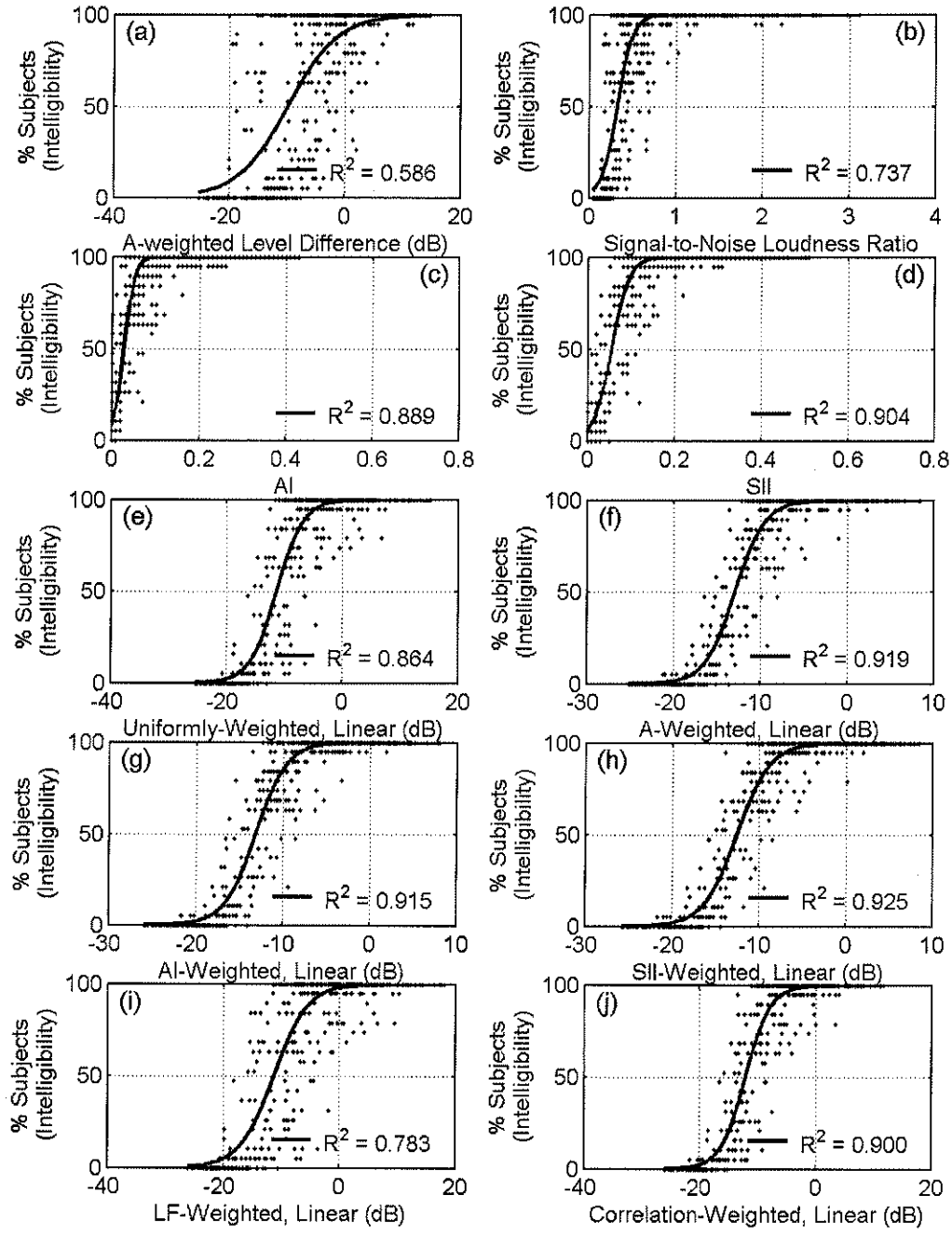


**Figure 13 Linear signal-to-noise ratio-summed measures: Threshold of Intelligibility: The dots are the fraction of the 19 subjects correctly identifying at least one word from each of the 340+160 sentences in the tests (500 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

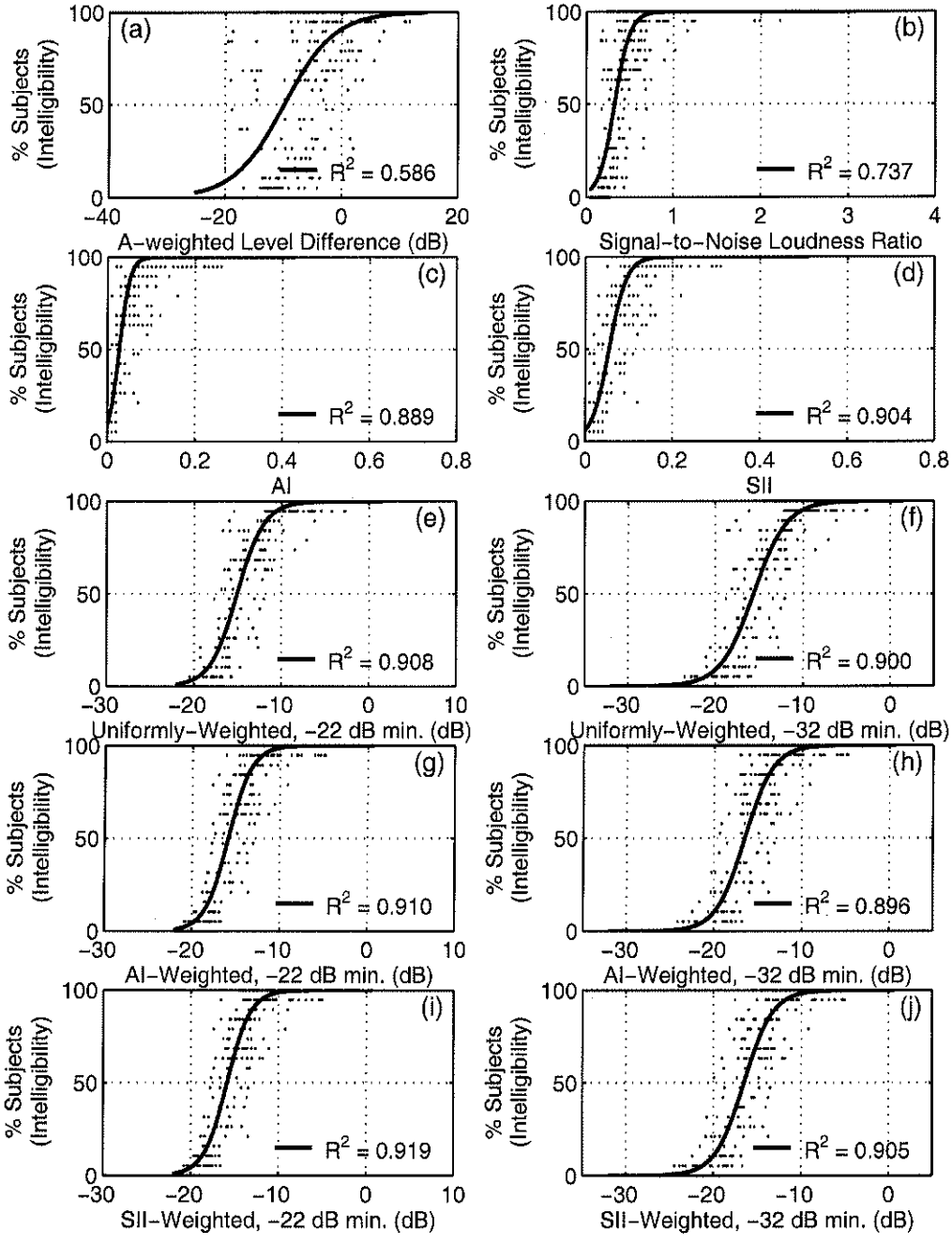


**Figure 14 Decibel level difference-summed measures: Threshold of Intelligibility: The dots are the fraction of the 19 subjects correctly identifying at least one word from each of the 340+160 sentences in the tests (500 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

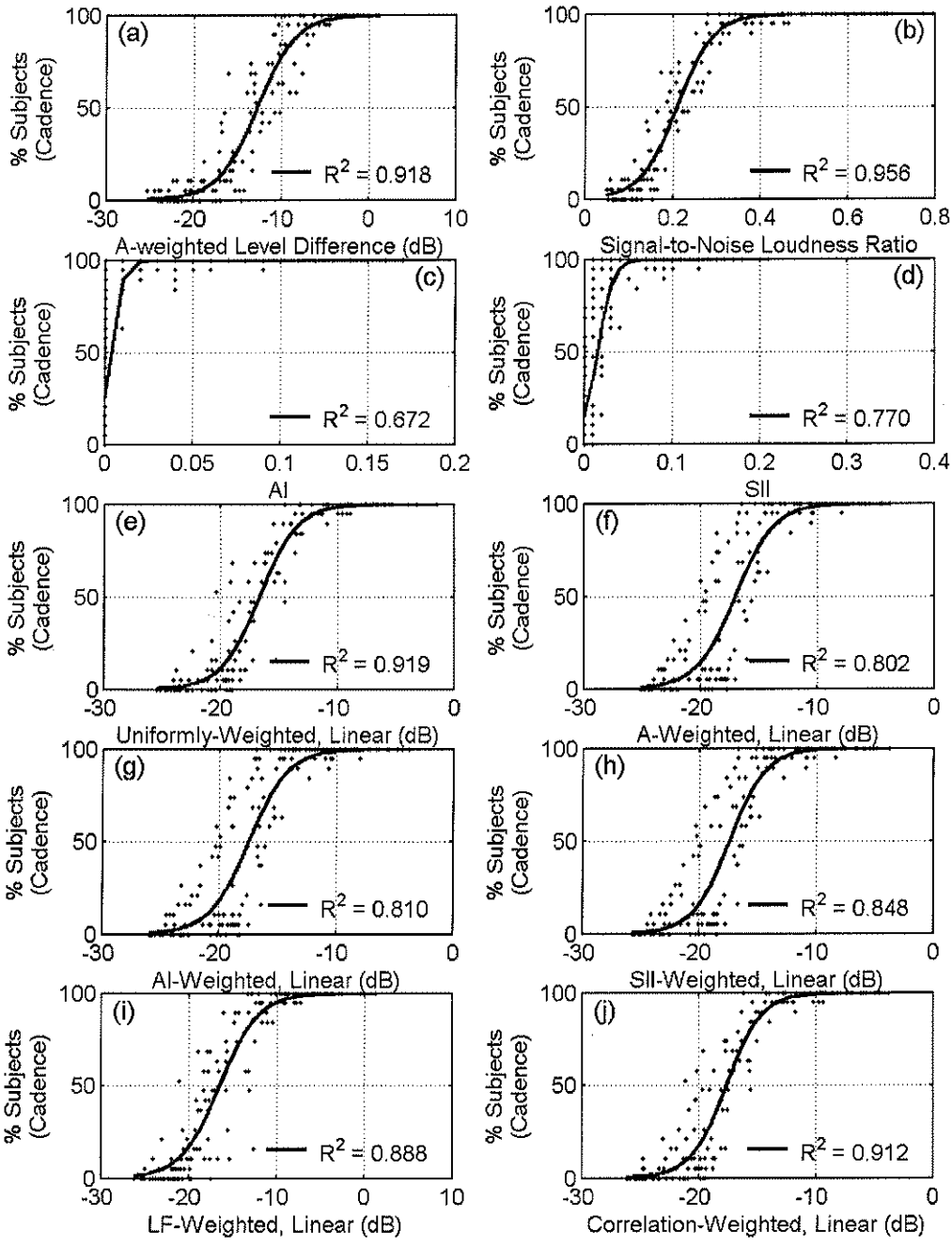


**Figure 15 Linear signal-to-noise ratio-summed measures: Threshold of Cadence: The dots are the fraction of the 19 subjects identifying the cadence of speech (including those correctly identifying some words) from each of the 160 sentences in the second test (160 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

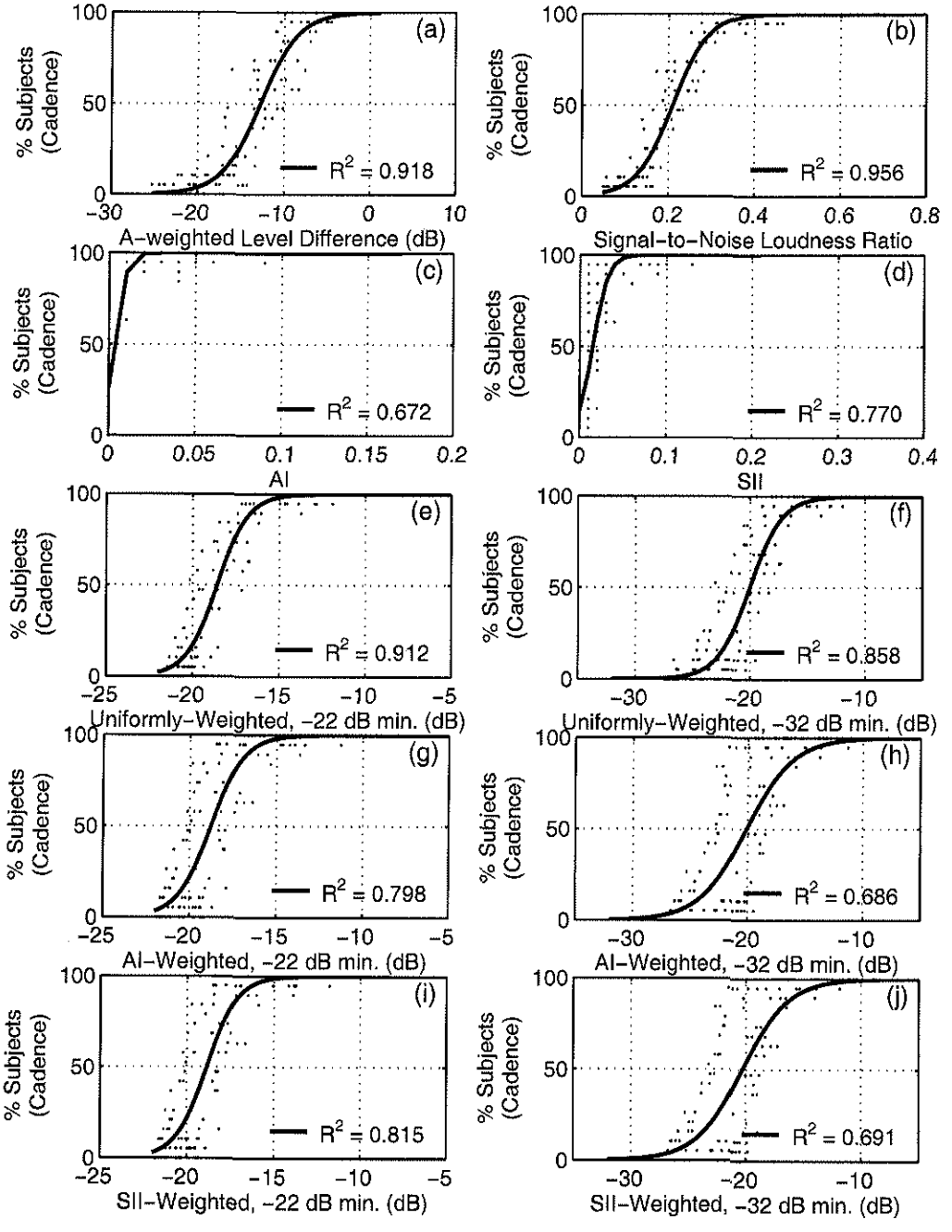


**Figure 16 Decibel level difference-summed measures: Threshold of Cadence: The dots are the fraction of the 19 subjects identifying the cadence of speech (including those correctly identifying some words) from each of the 160 sentences in the second test (160 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

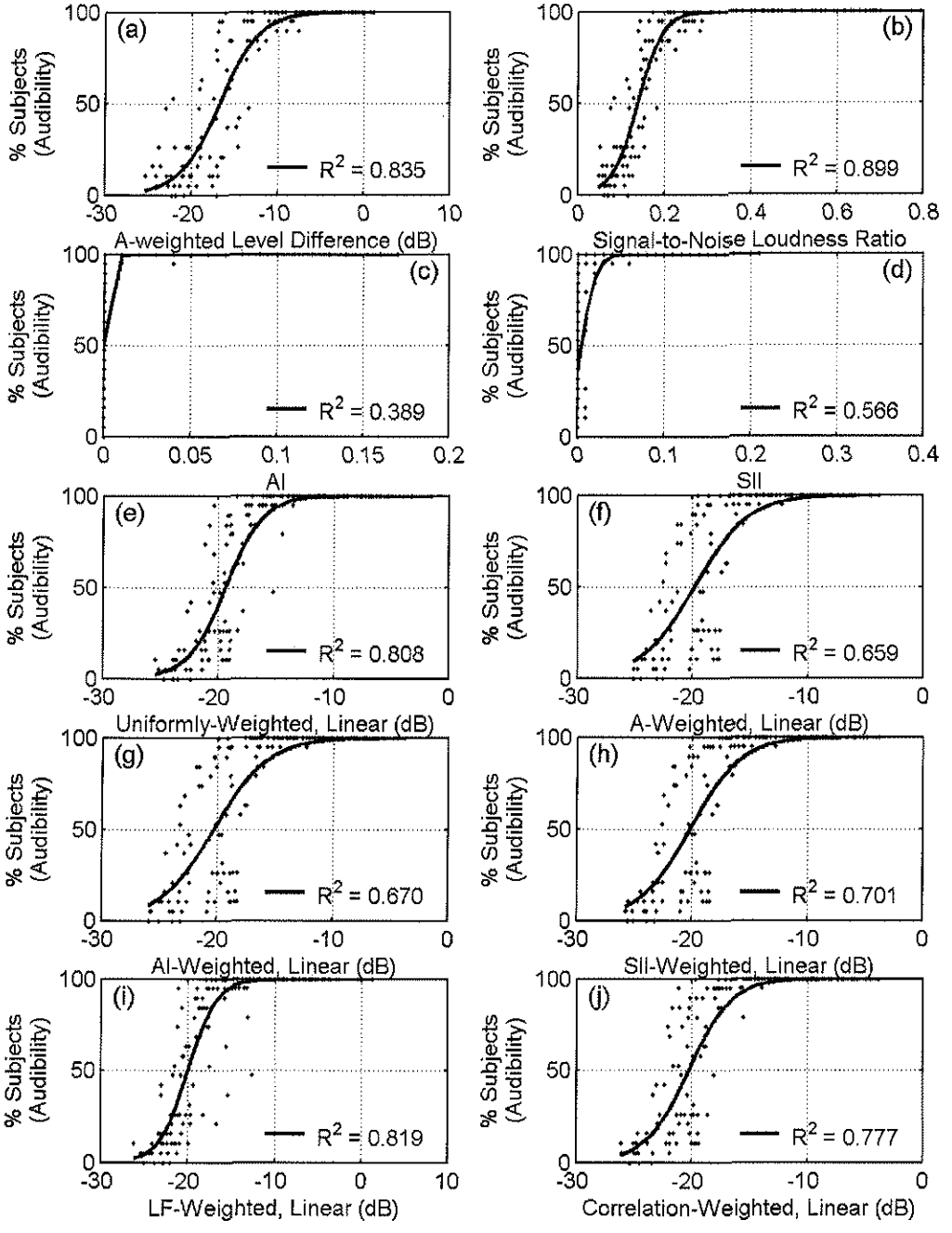


**Figure 17 Linear signal-to-noise ratio-summed measures: Threshold of Audibility: The dots are the fraction of the 19 subjects identifying the presence of speech (including those identifying cadence and those identifying some words) from each of the 160 sentences in the second test (160 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

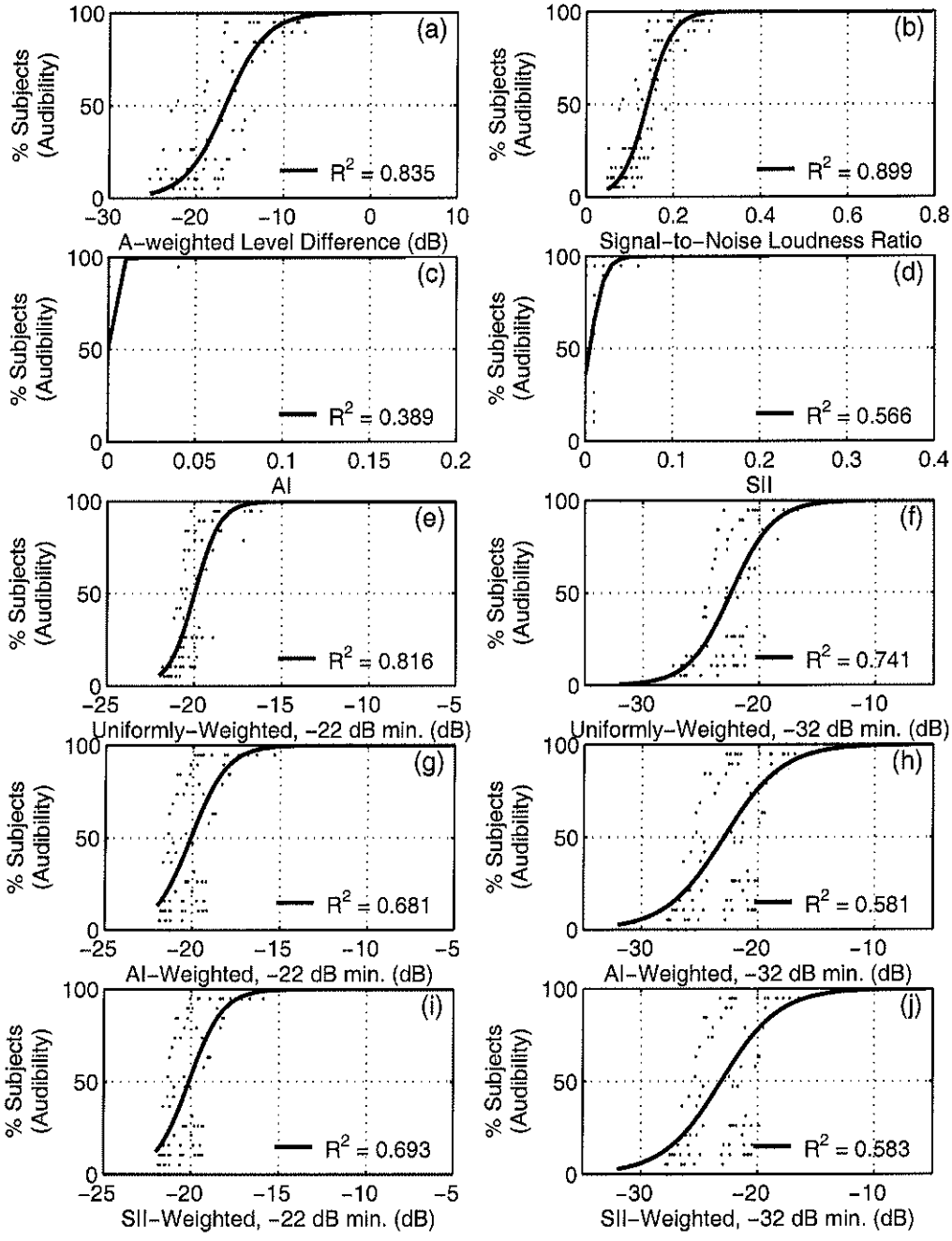


**Figure 18 Decibel level difference-summed measures: Threshold of Audibility: The dots are the fraction of the 19 subjects identifying the presence of speech (including those identifying cadence and those identifying some words) from each of the 160 sentences in the second test (160 dots total). The solid curve is the least-squares Boltzmann function fit to the mean. The squared correlation coefficient values ($R^2$) for this fit are shown.**

# 8 Discussion

The results of Sections 6 and 7 indicate that when the issue is intelligibility of overheard speech (as it often is), the weighted signal-to-noise type measures are appropriate, and in fact the SII-weightings seem best. When the issue is audibility (without understanding), A-weighted levels (or loudness ratios) are appropriate. This is reassuring to discover, given the origins of these measures. Weighted signal-to-noise measures were originally intended to relate to intelligibility; A-weighted levels are intended to relate to loudness, which is in turn an issue of audibility of sounds.

## *8.1 Intelligibility*

SII-weighted signal-to-noise level difference (dB), clipped to –22 dB is a good indicator of speech intelligibility for security situations. It has the highest $R^2$ (0.762) with intelligibility scores of all the measures considered (Figs. 11 and 12). Figure 19 shows the curve fits to the 5th, 25th, 50th, 75th, and 95th percentiles of the intelligibility score data collected from both listening tests, plotted versus SII-weighted, –22 dB clipped signal-to-noise ratio. Since the weights sum to unity, the minimum value of the measure is –22 dB, which is the endpoint of these curves. What the figure shows, for instance, is that at an index value of –15 dB, 75% of the subjects could identify 10% or more of the overheard words, 50% of the subjects could identify 23% or more of the words, and only the best 5% of the subjects could identify 92% or more of the words. Dropping an additional 5 dB to an index value of –20 dB, only 25% of the subjects could identify more than 4% of the words, and only the best 5% of subjects could identify 11% or more of the words.

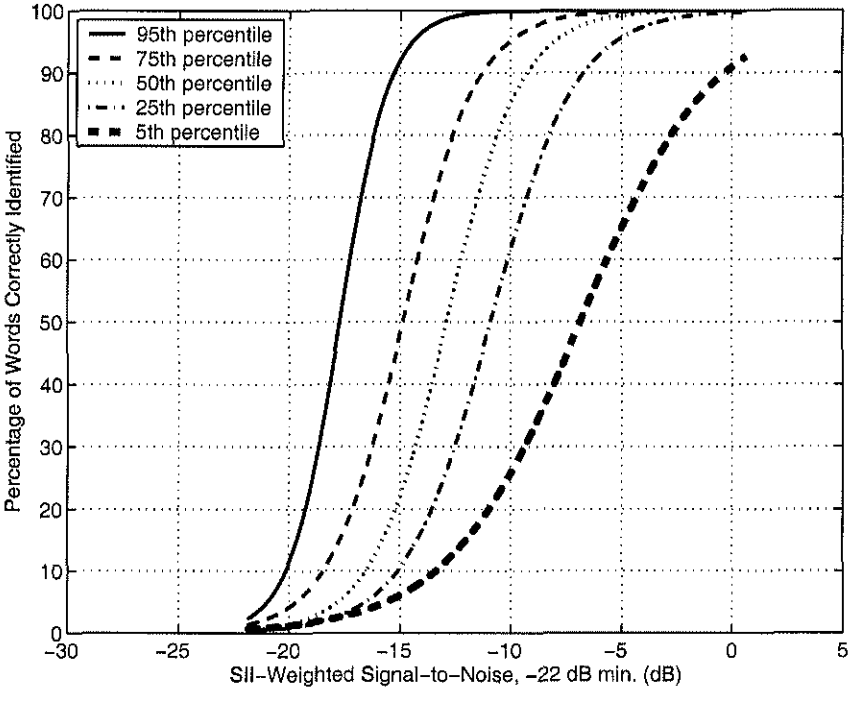


**Figure 19 Intelligibility score versus SII-weighted, dB-summed signal-to-noise. The curves from bottom to top represent the 5th, 25th, 50th, 75th, and 95th percentiles of the 9500 individual responses.**

SII-weighted (linear) signal-to-noise ratio is also a good indicator of speech intelligibility for security situations. It has the highest $R^2$ (0.744) with intelligibility scores of all the linear-summed measures considered (Fig. 11). Figure 20 shows the curve fits to the 5$^{th}$, 25$^{th}$, 50$^{th}$, 75$^{th}$, and 95$^{th}$ percentiles of the data collected from the listening tests. What the figure shows, for instance, is that at an index value of –10 dB, 75% of the subjects can identify 18% or more of the overheard words, whereas 50% of the subjects can identify 40% or more of the words, and only the best 5% of the subjects can identify 94% or more of the speech. Dropping an additional 5 dB of signal to an index value of –15 dB, only 25% of the subjects can identify more than 10% of the words, and only the best 5% of subjects can identify at least 40% of the words.

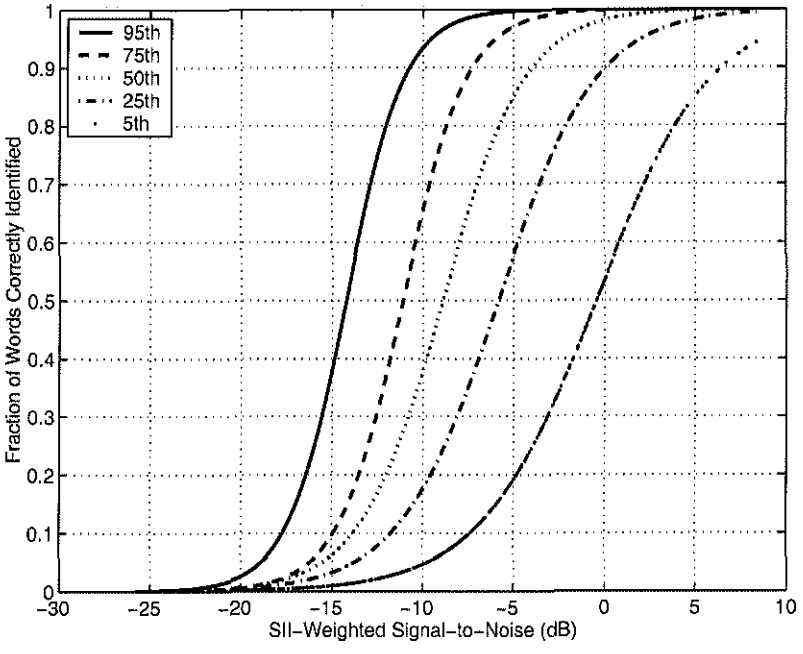


**Figure 20 Individual intelligibility score versus SII-weighted linear ratio-summed signal-to-noise ratio.**

It may not be important how many words can be understood by a listener: it could be argued that even one word is unacceptable, from a security point of view. This is why the Threshold of Intelligibility analysis is important. Figure 21 shows the intelligibility threshold curve for the percent of subjects able to correctly identify at least one word, for a given value of the –22 dB clipped SII-weighted signal-to-noise index. This measure had the highest $R^2$ (0.919) with intelligibility threshold of all the dB-summed measures considered (Fig. 14). For a measure value of –14 dB, 80% of the subjects can identify at least one word. 5 dB lower, at a value of –19 dB, only about 9% of subjects can identify at least one word. At an index value of –22 dB, less than 1% of subjects can identify a word.

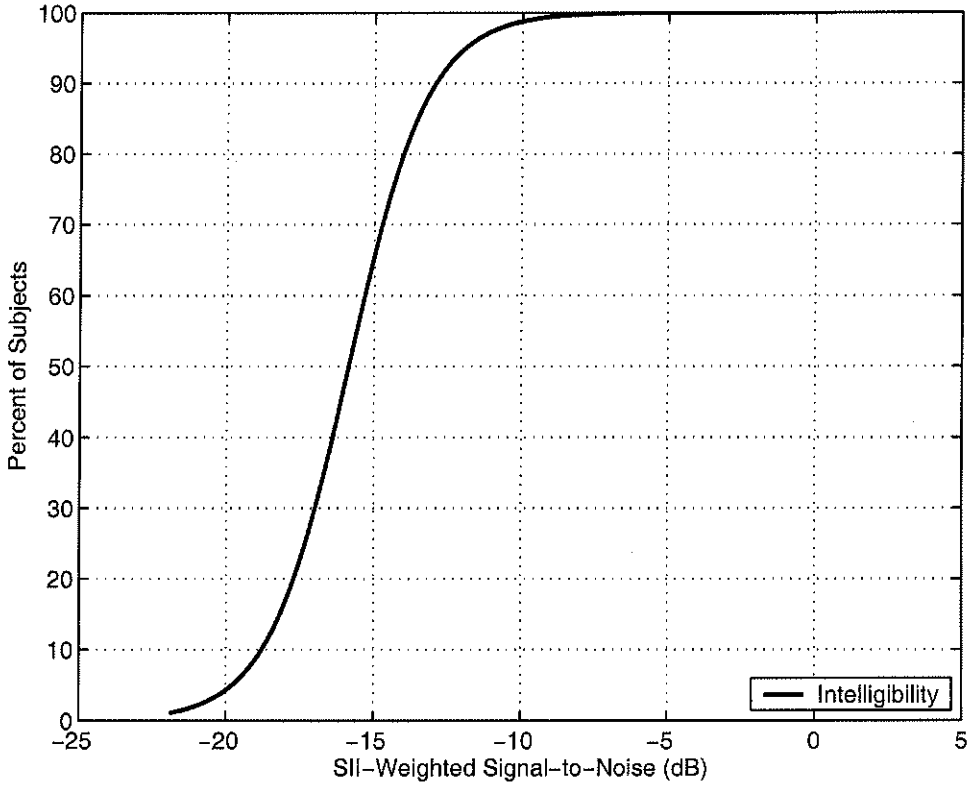


**Figure 21 Threshold of intelligibility versus –22 dB clipped SII-weighted signal-to-noise index. The curve represents the percentage of subjects able to correctly identify at least one word from a sentence for a given value of the index.**

Figure 22 shows the threshold curve for the percent of subjects able to correctly identify at least one word, for a given value of the SII-weighted linear ratio-summed signal-to-noise index. This measure had the highest $R^2$ (0.925) with intelligibility threshold of all the measures considered (Figs. 13 and 14).For a measure value of –10 dB, 80% of the subjects can identify at least one word. 5 dB lower, at a value of –15 dB, only 20% of subjects can identify at least one word. At an index value of –20 dB, less than 2% of subjects can identify a word.

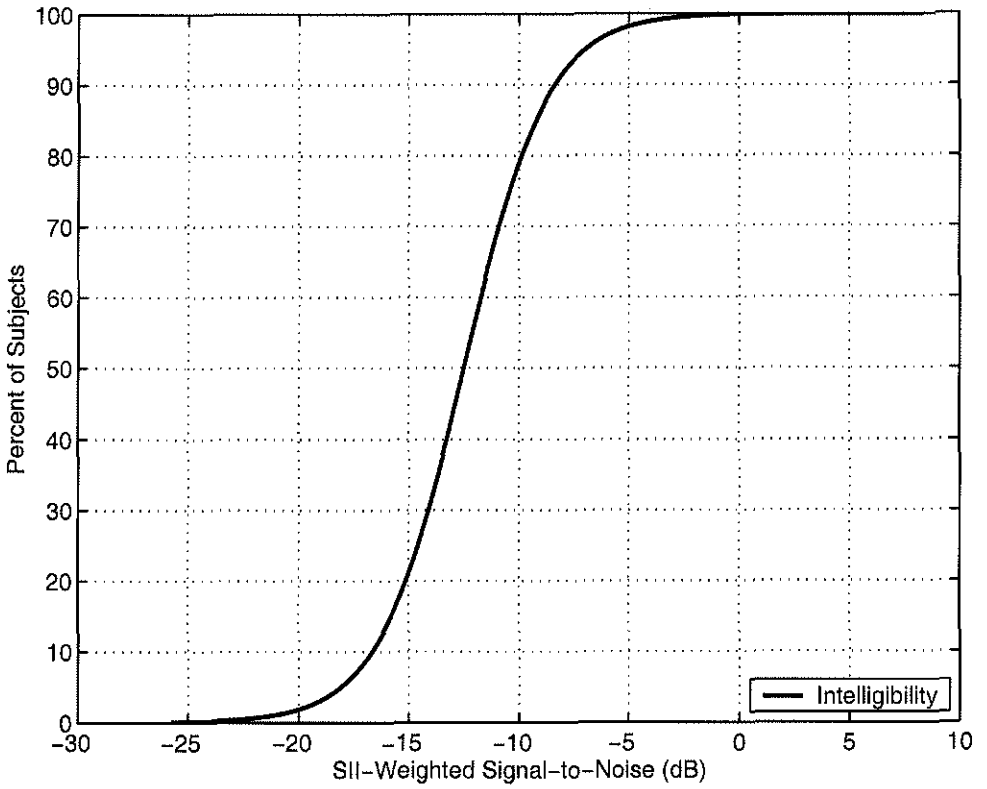


**Figure 22 Threshold of intelligibility versus SII-weighted linear-summed signal-to-noise index. The curve represents the percentage of subjects able to correctly identify at least one word from a sentence for a given value of SII-weighted signal-to-noise.**

## 8.2 Cadence and Audibility

Figure 23 shows curves for the threshold of cadence and the threshold of audibility: the percent of subjects able to detect the cadence, or the speech, versus A-weighted level difference. This measure had the second-highest $R^2$ (0.918 for cadence, 0.835 for audibility) of all the measures considered (second only to loudness ratio, which is less conventional and more difficult to compute). At –10 dB (i.e., speech level 10 dB *lower* than background noise level), 95% of the subjects can tell that there is a sound other than the background noise, and 80% of the subjects can identify the cadence of speech. 5 dB lower than this, at a level difference of –15 dB, 67% of subjects can hear the sound, 27% identifying the cadence. The threshold of intelligibility curve was not included in Figure 23 since the correlation is so poor ($R^2$ = 0.586). A-weighted level difference is *not* a good measure for assessment of threshold of intelligibility.

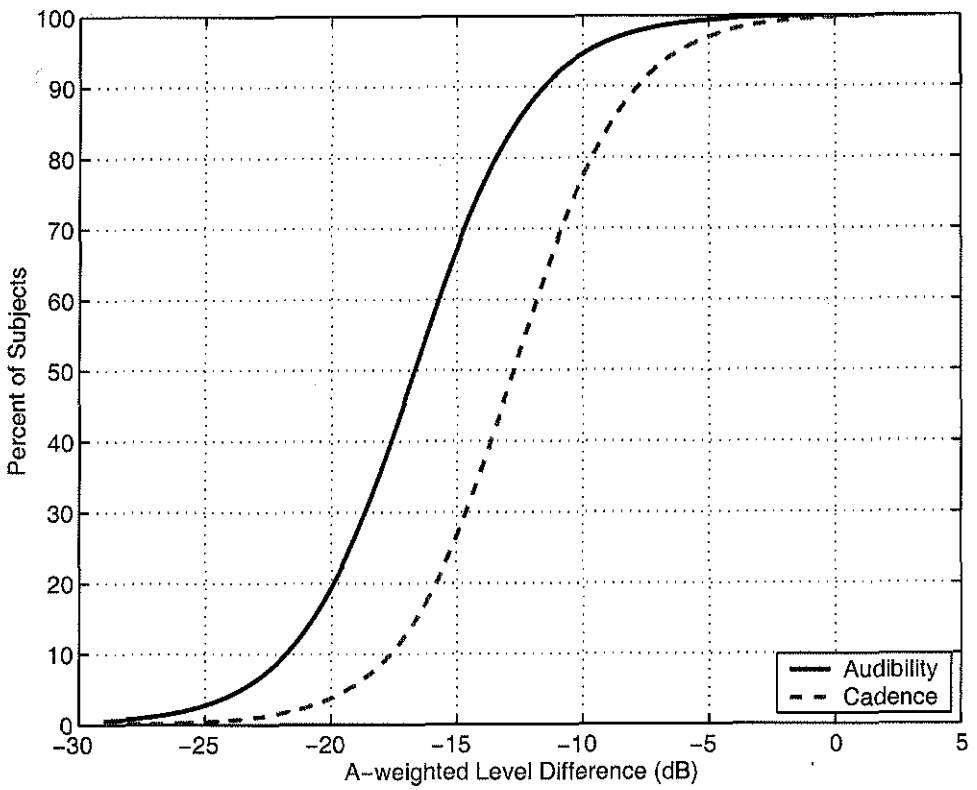


**Figure 23 Thresholds of cadence and audibility versus A-weighted level difference. The curves represent the percentage of subjects able to at least identify the cadence of speech (dashed line) or to at least detect the presence of speech (solid line) for a given value of A-weighted Level Difference.**

The weighted signal-to-noise-type measure that is best correlated with the thresholds of cadence and audibility is the uniformly-weighted –22 dB clipped signal-to-noise. However, as discussed in Sec. 7, the fit curves do not reach zero over the range of the measure (i.e., at the minimum value of –22 dB). Therefore, the best-correlated useful measure is the –32 dB clipped, uniformly weighted signal-to-noise ratio index (see Figs. 15–18). In fact, this measure is also well-correlated with the threshold of intelligibility (see Fig. 14). The uniformly-weighted –32 dB clipped signal-to-noise can therefore be used to assess all three thresholds, and, importantly, indicate relative relations among them. The usefulness of uniformly-weighted band levels has been explored by Tachibana et al. as an indicator of loudness [25]. This ties in well with the above-noted observation that detection of the thresholds of audibility and of cadence has to do with the loudness of the speech sounds.

Figure 24 shows the threshold curves for the percentage of subjects able to correctly identify at least one word (Intelligibility), able to identify at least the cadence (Cadence), and able to at least hear some speech sounds (Audibility) versus uniformly-weighted, –32 dB clipped signal-to-noise index. For a measure value of –15 dB, 98% of the subjects could hear some speech sounds, and 60% could identify at least one word. 5 dB lower, at

a value of –20 dB, only 8% of subjects could identify at least one word. At an index value of –25 dB, less than 1% of subjects could identify a word, and only 20% could hear speech sounds. If one considers 50% of the subjects as a threshold point (as is the norm), then the threshold of intelligibility could be said to be at about –15.5 dB, the threshold of audibility about 7 dB lower, at –22.5 dB.

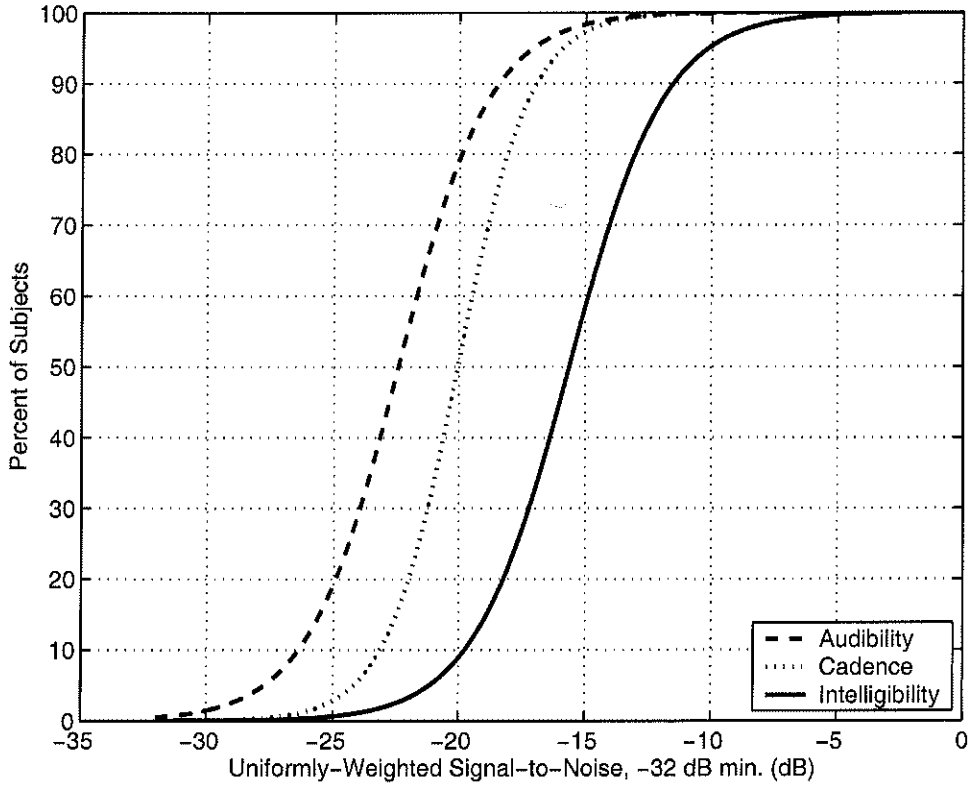


**Figure 24 Thresholds versus uniformly-weighted, –32 dB clipped signal-to-noise index. Each curve represents the percentage of subjects able to correctly identify: at least one word from a sentence (Intelligibility), the cadence of speech (Cadence), or the presence of speech sounds (Audibility).**

## 8.3 Subject Group

The group of participants used in the tests consisted of people with excellent hearing. There are the issues of what percentage of the general population this group represents, and what these results imply for the population at large. The former is more easily addressed than the latter. There is an ISO standard relating statistics of hearing loss for "otologically normal" listeners, broken down by age and gender [18]. From the data provided therein, an "average" otologically normal listener can be constructed by averaging the HL data for 30 and 40 year old males and females. This has been done, and the $10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, and $90^{th}$ percentiles of the HL curves are shown in Figure 25. Also shown in the figure is the average HL curve for the 19 "better" subjects used in the analyses in this report. Observe that the shape of the subject curve deviates from the percentiles above 2 kHz or so; it is possible that our subjects were not all "otologically

normal", or this may simply be a result of smoothing of the data in the ISO standard. Nevertheless, this figure, together with the discussion of F-HL above in Section 4.7, lends evidence to support the assertion that our subjects had excellent hearing, but not "unreasonably" so. 25% of the population of this average 30/40-year old male/female listener can be expected to have hearing as good or better.

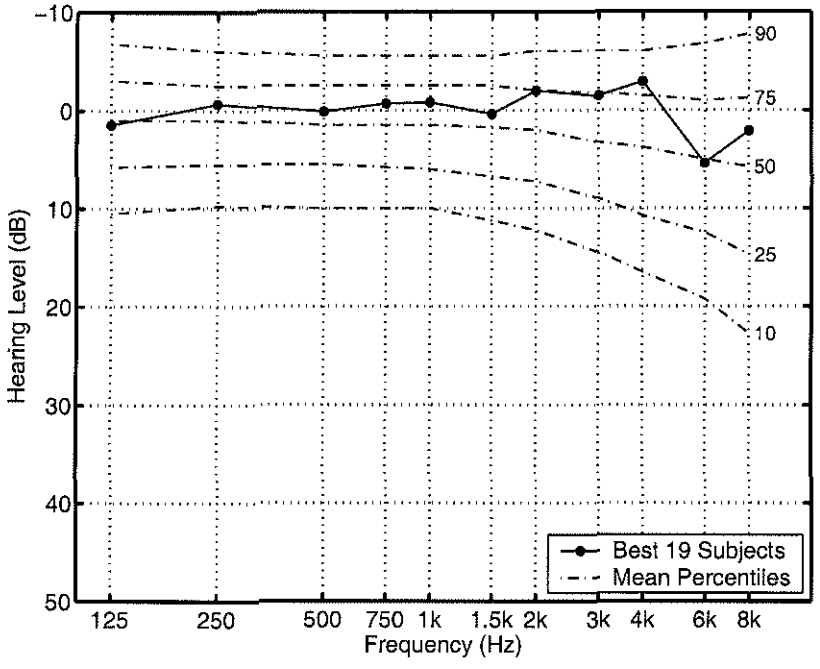


**Figure 25 HL curve for best 19 subjects shown with 10th, 25th, 50th, 75th, and 90th percentile HL curves for an "average" otologically normal listener. The percentiles give the percentage of the population with hearing loss *worse* than the curve. The "average" listener response is the mean of the 30-year old and 40-year old male and female responses given in ISO 7029-1984.**

Knowing what portion of the population was used as test subjects is important, but does not necessarily enable "extrapolation" of the results to groups with worse hearing. For certain, people with worse hearing will get lower scores on the listening tests (and therefore be less of a risk, from a security point of view). However, there are factors other than hearing loss which can cause reduced scores. Figure 26 shows plots of individual intelligibility score versus F-HL for the subjects in this study (dots), and for a similar study conducted for open-office situations (crosses) [24]. The average scores were higher for the open-office study, since the test was "easier", but a similar trend is seen for both. The mean score does indeed drop with increasing hearing loss, but the $R^2$ values of the fits are rather low (0.31 and 0.61). This can be interpreted as meaning that there is scatter in the data not attributable to the measured hearing loss. As mentioned above in Section 4.7, there are other factors involved.

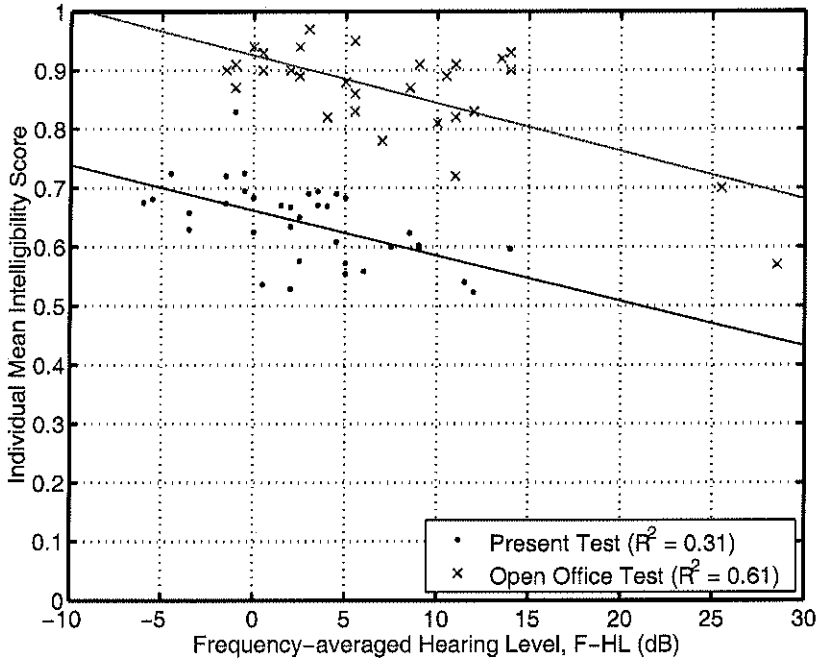


**Figure 26 Intelligibility test scores versus frequency-averaged hearing level for both the present test (dots) and a similar test conducted for open-office workstation privacy (crosses).**

# 9 Conclusions

Listening tests simulating speech transmission through a range of typical office wall constructions have been used to find objective measures of intelligibility and audibility suitable for architectural speech security situations. It has been found that a weighted speech-signal-to-noise ratio, using the band importance frequency weights from the SII calculation, is a good measure of speech intelligibility, and the threshold of intelligibility. It has also been found that, below the threshold of intelligibility, simple A-weighted level difference is able to accurately predict audibility of the speech or its cadence. Furthermore, a uniformly-weighted 1/3-octave-band signal-to-noise ratio (with band level differences clipped to –32 dB) is a good indicator for all three thresholds

Existing measures such as AI and SII are not suitable for assessing speech security. While they are highly-correlated with the listening test intelligibility scores, both fail to indicate zero intelligibility at their minimum values of zero. Furthermore, they provide no information regarding the thresholds of audibility and cadence.

Loudness is as good a predictor of audibility of speech or its cadence as the A-weighted levels, but is more difficult to compute. Neither is well-correlated with intelligibility scores, so neither is suitable for defining the threshold of intelligibility.

The relationships derived in this report between intelligibility scores and thresholds and the objective measures are "worst-case" from a security point of view. They are therefore broadly applicable. The listeners were acute-hearing and cued to expect to overhear speech, the speaking voice was strong and clear. Predicting a level of security from this work should err on the conservative side in the case of listeners with poorer hearing and/or talkers with less clear voices.

## Acknowledgements

The authors gratefully acknowledge the assistance of Ms. Kimberlee Cuthbert, who was responsible for scheduling subjects, running subject experiments, and assembling the measurement data, and Mr. Ketana Teav, who set up much of the equipment and wrote control software for the computer and hardware. We also would like to thank the subjects who each volunteered several hours of their time to help us do this work.

Funding for this work provided by Public Works and Government Services Canada (PWGSC) and the Royal Canadian Mounted Police (RCMP).

# Appendix A: Pilot Study Summaries

As discussed above in Section 3, settling on a final test procedure involved the assessment of a number of variables. Several pilot tests were conducted to address the importance of many of these variables.

## *Subject Response and Scoring*

The initial question was to decide on subject response and scoring process. One possible approach was to have the subjects listen to continuous speech discourse and instruct them to respond when a given level of security is established. The sound field could be established by letting the subjects adjust speech and/or noise levels manually, or by having the level(s) change automatically, stopping when the subjects respond. This latter is the approach used in Cavanaugh et al. [5]. What this approach lacks is a quantifiable intelligibility score, such as can be obtained through the use of test sentences. Test sentences can be scored in terms of the number of words correctly understood, and thereby resulting in an unequivocal measure of the intelligibility (i.e., of the speech security).

Initial informal tests established the oral response and scoring procedure described in Section 4, and revealed the need to expose subjects to a test sentence only once. If a subject is presented a sentence for a second time, it is quite likely that words can be correctly understood from memory, even if the sentence is not entirely intelligible.

## *Multiple Sentences*

Another factor which became apparent through the initial testing was that different sentences played back with the same level, filtered by the same wall, and paired with the same background noise can be of different intelligibility. The reason, it would seem, is that each sentence has a slightly different spectrum. The "interaction" of the source speech with the wall TL resulted in a "transmitted" speech spectrum which was unique for each sentence, even if the "physical" parameters (speech level, wall, background noise level and spectrum) were constant. In these low-intelligibility cases, a slight variation in level (or in SII-weighted signal-to-noise, as it were) can lead to large variations in intelligibility. Some sentences are just "harder" to understand than others. In a pilot test, 10 subjects each listened to 10 sentences for each of 6 "difficult" conditions (where all subjects neither correctly understood all the words nor zero of the words). On average, the standard deviation of responses for one subject for one "condition", calculated from all 10 sentences, was about 30%.

## *Absolute Level*

It was pointed out by Cavanaugh et al. [5] that "speech privacy is related to speech intelligibility rather than to level." In intelligibility theories, the assumption is as long as the sounds are neither so faint as to be near the threshold of hearing nor so loud as to be near the threshold of feeling, the absolute level is not important. Two speech-in-noise conditions with a constant level *ratio* but different absolute level are seen as equivalent. To explore this assertion, a pilot test was conducted. 10 subjects each listened to 10

sentences at each of 3 different levels, but at the same A-weighted level ratio, for two different wall types. There was no clear trend relating intelligibility score to absolute level. For one of the wall types, the three speech levels were 38 dBA, 33 dBA, and 28 dBA. For the second wall type, the speech levels were 40 dBA, 35 dBA, and 30 dBA. In the first instance, the mean score for the 38 dBA level was just statistically different than the 33 dBA level, but not from the 28 dBA level; the latter two were not different from each other either. The results for the second wall were similar. For the testing, it was therefore decided not to attempt to control the overall level of the test sentences, except to keep away from the threshold of hearing, and to keep below 55 dBA noise level.

### *Different Talkers*

As in the case of different sentences spoken by the same talker, the same sentence spoken by different talkers—or even by the same talker with different effort—will have a different spectral energy distribution. A pilot test was conducted repeating a 10-sentence test for seven different recordings of the test sentences (3 female talkers, 4 male talkers). The wall TL, noise spectrum, and noise level was the same for all of the sentences. The 3 female voices were set 3.5 dBA quieter than the male voices, as is correct for equivalent effort [15]. 70 unique sentences were used (meaning each talker did not speak the same words), to avoid subject familiarity with the material. The voice that was most easily heard by the subjects was a male, speaking clearly, and also happened to be the highest-quality recording (digitally speaking: 16-bit, 44.1 kHz, as opposed to 16-bit, 16 kHz for the others). This is the voice used in the main tests.

### *Noise On/Off*

The comment is sometimes made that persistent background noise is not really noticeable until it stops. This is observed oftentimes in offices when the air-handling shuts off at the end of the work day. To investigate whether this "acclimatization" has an effect on the intelligibility tests, a pilot test was conducted. Subjects conducted a 20-sentence test as described above in Section 3, and an additional test using the same walls, noises, levels, and speaking voice (but 20 different sentences), but with the noise not turned off between sentences. There was no significant difference found between the scores on the two tests. The intelligibility scores for the "noise on/off" test were slightly higher (i.e., the security was slightly worse). This is presumed to be because the subjects were "cued" by the onset of the noise to prepare themselves for the task of identifying the words. When this cue was absent (when the noise persisted), the subjects were sometimes "surprised" by the onset of speech, which was a comment made by several of the subjects.

## Appendix B: Intelligibility Scoring

As discussed above in Section 4.5, there are alternative views on how to score the sentences for the intelligibility testing [17]. The distinction is made between "key" words (nouns, verbs, adjectives) and "connective" words, with the rationale being that "connective" words do not necessarily carry the meaning of the sentence. Given a major difference, it would be up to a security designer to decide whether it is the "words" or the "meaning" that is to be kept secure. This is a complicated issue, so without getting into the details, the relationships among the scoring strategies are simply presented here. Figure 27 shows the relationship between (a) Full score and Keyword score; (b) Full score and Weighted score; and (c) Keyword score and Weighted score. Notice that all scoring strategies are highly-correlated with each other.

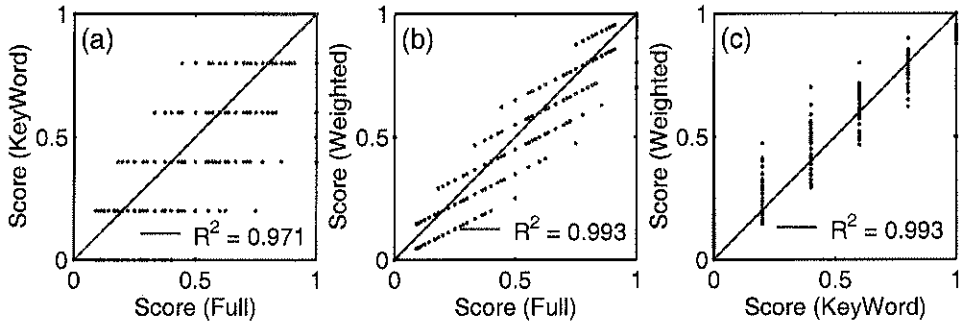


**Figure 27 Relationships among different scoring strategies. The dots are the 9500 subject responses from the listening tests. The line has slope 1.**

Figure 28 is analogous to Figure 11 showing the individual subject intelligibility test scores versus various indices, for the Keyword scores. Figure 29 is similar for the Weighted scores. Notice that the rank-ordering of the squared correlation coefficients is the same as for the full scores: AI-weighted and SII-weighted signal-to-noise highest, A-weighted level difference and loudness ratio lowest. The details of the relationship between the scores and the measures are, however, slightly different, which conceivably could lead to different criteria or ratings for security for a given value of the objective measure.

Figure 30 shows the threshold of intelligibility curve shown in Figure 22 for the Full scores, along with the similar curves derived from considering Keyword and Weighted scores, all plotted versus SII-weighted signal-to-noise. There is no difference between the Full and Weighted scores, since scoring half a point for understanding a connective word is the same a scoring a full point—a word was correctly overheard and the intelligibility score was greater than zero. Above –15 dB, however, the Keyword score curve is slightly to the right, indicating a "looser" security condition. For a given index value, a smaller proportion of subjects are able to hear the "key" words than the "full" words.

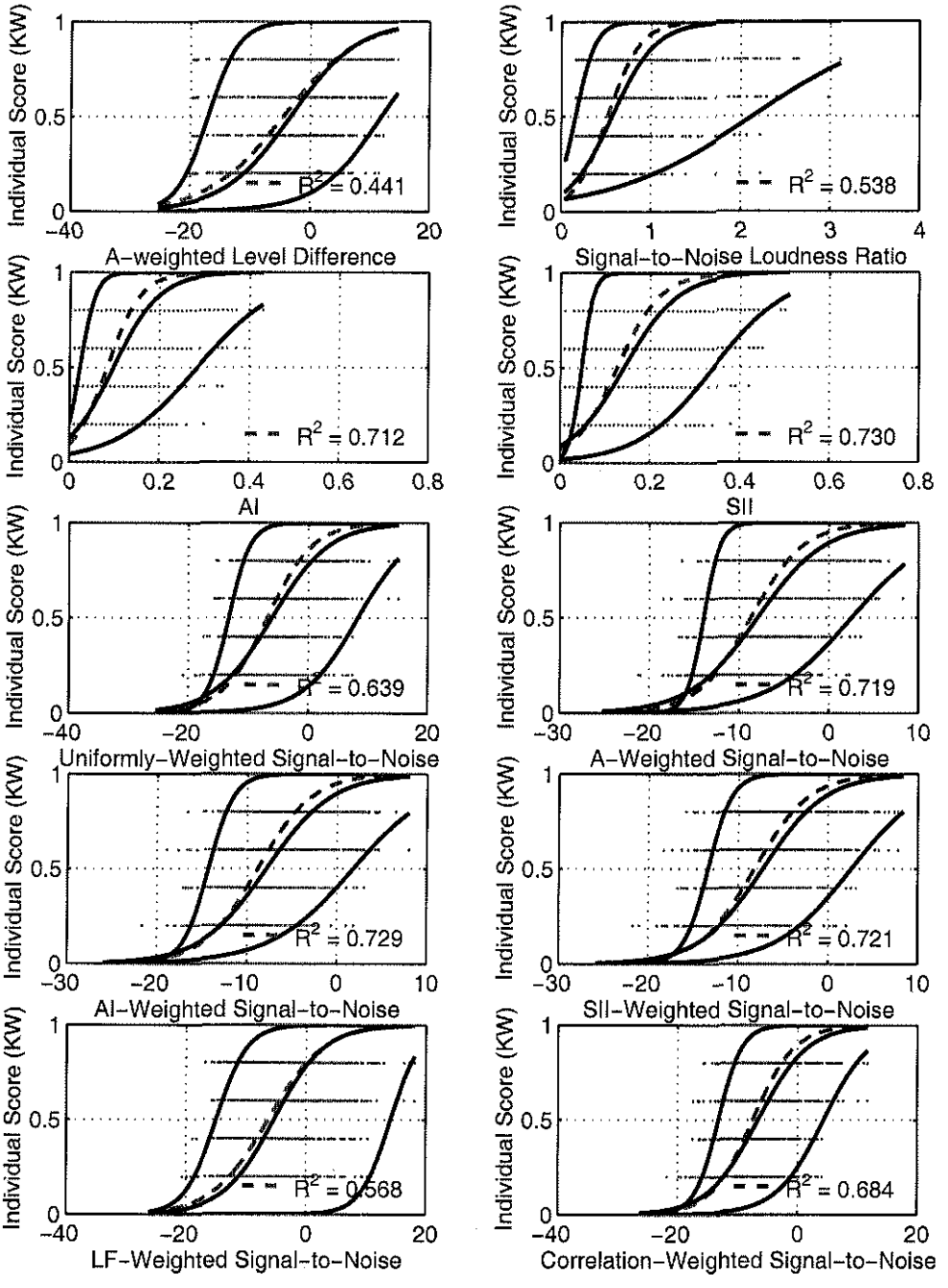


**Figure 28 Keyword intelligibility scores (as a fraction from 0 to 1) for 19 subjects on 500 sentences each (9500 points total), plotted versus various linear signal-to-noise ratio-summed indices.**

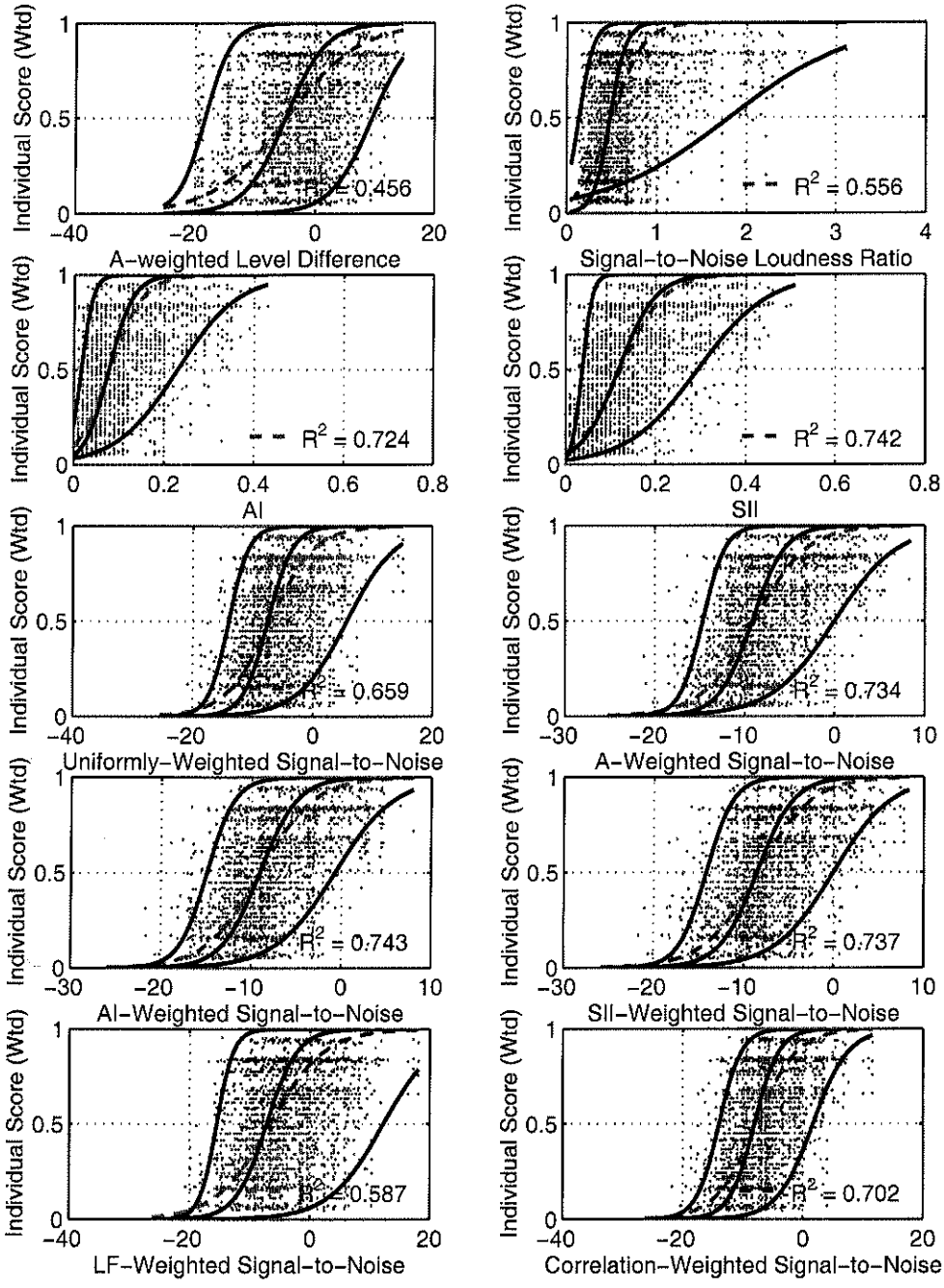


**Figure 29 Weighted intelligibility scores (as a fraction from 0 to 1) for 19 subjects on 500 sentences each (9500 points total), plotted versus various linear signal-to-noise ratio-summed indices.**

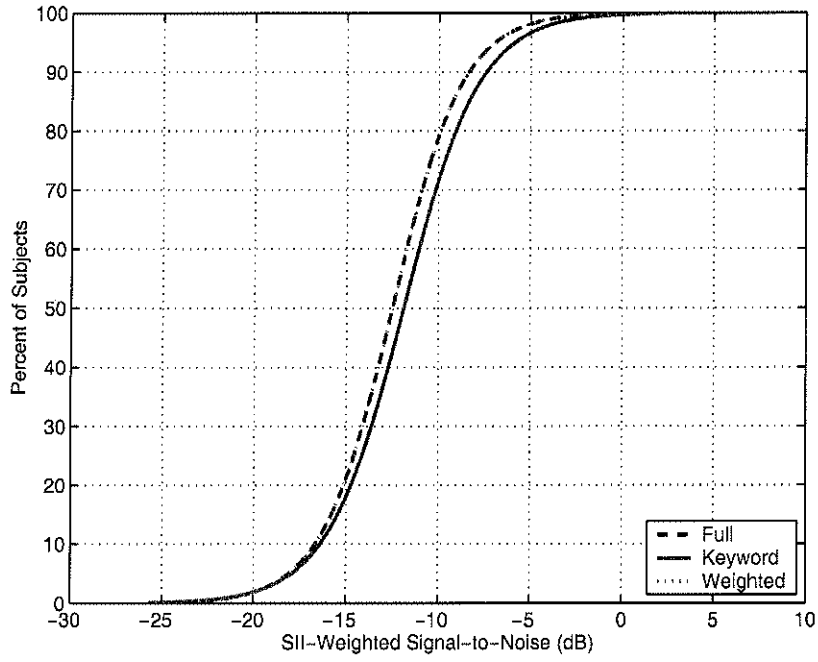


**Figure 30 Threshold of intelligibility versus SII-weighted linear ratio-summed signal-to-noise index for three scoring strategies: Full Score (dashed curve), Keyword Score (solid curve), and Weighted Score (dotted line). The Full and Weighted curves lie exactly on top of each other since the threshold where scores become non-zero is the same for both—it doesn't matter whether a connective word is worth a full or half point, it's still non-zero.**

## Appendix C: Signal-to-Noise Clipping Level

The parameter $L$ introduced in Sec. 5.4 defines the minimum value of the signal-to-noise level difference that is taken as contributing to the measure $X_w$. The results of the listening tests were correlated with $X_w$ for $L = -12$ dB to $L = -32$ dB, in 2 dB steps for the three weighting strategies (uniform, AI, and SII). Figure 31 shows the resulting $R^2$ values for the relationships of the measures with each of: (a) intelligibility scores, (b) threshold of intelligibility, (c) threshold of cadence, and (d) threshold of audibility. Clipping below –32 dB does not appreciably change the relationships; little of the collected data lies in that range.

For the intelligibility scores and the threshold of intelligibility, the correlations are much the same for all three weightings; the SII only slightly higher (notice the scale on the ordinate axis). There is a peak at around –22 dB, which is also about the point where the fit curve trends to zero intelligibility (see Figs. 8 and 10). Clipping below this, the $R^2$ trends to a constant value. Above this, not only does the correlation drop, but also the data starts to clip before the intelligibility reaches zero, as happens with AI and SII.

The thresholds of cadence and audibility show that uniform weighting is better correlated than the AI and SII for all clipping levels. The $R^2$ actually drops with decreasing clipping level, but from inspection of Figure 16 and Figure 18, a level of –32 dB is necessary to ensure the data points are not clipped before audibility reaches zero.

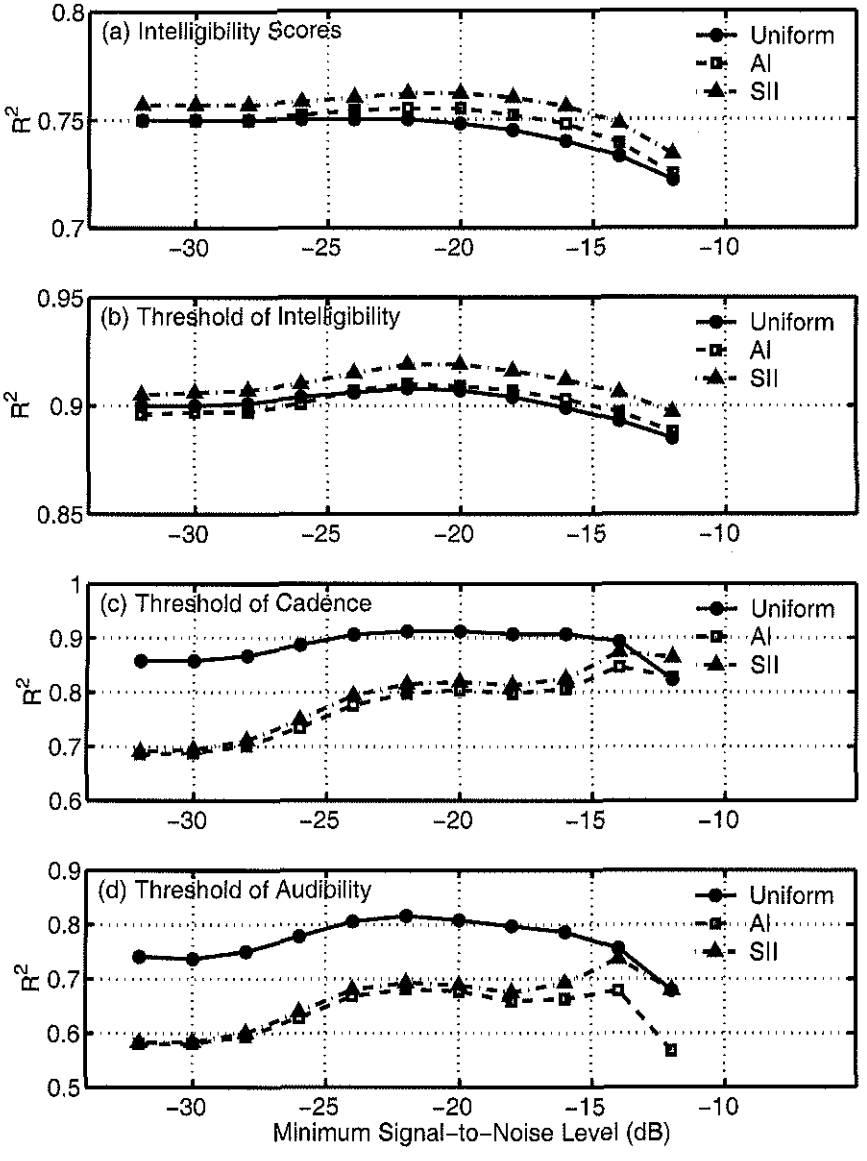


**Figure 31 Squared correlation coefficient (R2) versus minimum band signal-to-noise clipping level. The correlation describes the relationship between the subjective test responses and the measure calculated using the corresponding level, for each of three weightings (uniform, AI, and SII) for (a) intelligibility scores, (b) threshold of intelligibility, (c) thresholds of cadence, and (d) threshold of audibility.**

## References


[1] ASTM E1130-02, "Standard Test Method for Objective Measurement of Speech Privacy in Open Offices Using Articulation Index," ASTM International, West Conshohocken, PA, U.S.A.

[2] ANSI S3.5-1969, "American National Standard Methods for the Calculation of the Articulation Index," ANSI, New York, NY, U.S.A.

[3] ANSI S3.5-1997, "American National Standard Methods for Calculation of the Speech Intelligibility Index," ANSI, New York, NY, U.S.A.

[4] ASTM E336-97, "Standard Test Method for Measurement of Airborne Sound Insulation in Buildings," ASTM International, West Conshohocken, PA, U.S.A.

[5] W.J. Cavanaugh, W.R. Farrell, P.W. Hirtle, and B.G. Watters, "Speech Privacy in Buildings," J. Acoust. Soc. Am., **34**, 475–492 (1962).

[6] R.W. Young, "Re-Vision of the Speech-Privacy Calculation," J. Acoust. Soc. Am., **38**, 524–530 (1965).

[7] B.R. Wyerman, "A study of speech privacy in closed offices," J. Acoust. Soc. Am., **72**, 436–442 (1982).

[8] K.P. Roy, "Speech privacy between closed and open offices," Presented at the 136th Meeting of the Acoustical Society of America (1998).

[9] M. Vorländer and R. Thaden, "Investigation of speech privacy in buildings by means of auralisation," Proc. 2nd FIA, SEA, Spanish Acoustical Society (2000).

[10] D.J. Caswell, Y. Ma, J.T. Goodchild, and L. Dai, "A study on speech privacy in closed offices," Presented at the 141st Meeting of the Acoustical Society of America (2001).

[11] A.C.C. Warnock, "Acoustical privacy in the landscaped office," J. Acoust. Soc. Am., **53**, 1535-1543 (1973).

[12] C. Speaks, B. Parker, C. Harris, and P. Kuhl, "Intelligibility of connected discourse," J. Speech and Hearing Research, **15**, 590-602 (1972).

[13] F. Santon, "Intelligibilité de la parole masquée et rôle de la direction de la source masquante," Acustica, **61**, 67–74 (1986).

[14] "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio and Electroacoustics, **17**, 227–246 (1969).

[15] W.O. Olsen, "Average speech levels and spectra in various speaking/listening conditions: A summary of the Pearson [sic], Bennett, & Fidell (1977) report," Am. J. of Audiology, **7**, 1–5 (1998).

[16] ANSI S12.2-1995 (R1999), "American National Standard Criteria for Evaluating Room Noise," ANSI, New York, NY, U.S.A.

[17] J.P. Egan, "Articulation testing methods," Laryngoscope, **58**, 955–991 (1948).

[18] ISO 7029-1984 (E), "Acoustics — Threshold of hearing by air conduction as a function of age and sex for otologically normal persons," ISO, Geneva, Switzerland.

[19] ANSI S1.8-1989 (R2001), "American National Standard Reference Quantities for Acoustical Levels," ANSI, New York, NY, U.S.A.

[20] ISO 532-1975 "Acoustics -- Method for calculating loudness level," ISO, Geneva, Switzerland.

[21] E. Zwicker, H. Fastl, and C. Dallmayr, "BASIC-Program for calculating the loudness of sounds from their 1/3-oct band spectra according to ISO 532 B," Acustica, **55**, 63–67 (1984).

[22] "Origin User's Manual Version 5," Microcal Software Inc., 1997.

[23] R.J. Carroll and D. Ruppert, Transformation and Weighting in Regression, (Chapman and Hall, New York, 1988).

[24] J.S. Bradley and B.N. Gover, "Describing Levels of Speech Privacy in Open-Plan Offices," IRC Research Report 138, 12 September 2003, available from http://irc.nrc-cnrc.gc.ca/ircpubs

[25] H. Tachibana, Y. Sonoda, K. Iwamoto, S. Kuwano, and S. Namba, "Validity of arithmetic average of sound pressure levels in octave bands as a loudness index," J. Acoust. Soc. Jpn. (E), **14**, 197–204 (1993).