International Journal of Occupational Safety and Ergonomics (*JOSE*) 2004, Vol. 10, No. 2, 179–189

Speech Intelligibility of the Callsign Acquisition Test in a Quiet Environment

Misty Blue
Celestine A. Ntuen

Institute for Human-Machine Studies, Department of Industrial & Systems Engineering, North Carolina A&T State University, Greensboro, USA

Tomasz Letowski

Human Research & Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Grounds, MD, USA

*This paper reports on preliminary experiments aimed at standardizing speech intelligibility of military Callsign Acquisition Test (CAT) using average power levels of callsign items measured by the Root Mean Square (RMS) and maximum power levels of callsign items (Peak). The results obtained indicate that at a minimum sound pressure level (*SPL*) of 10.57 dBHL, the CAT tests were more difficult than NU-6 (Northwestern University, Auditory Test No. 6) and CID-W22 (Central Institute for the Deaf, Test W-22). At the maximum* SPL *values, the CAT tests reveal more intelligibility than NU-6 and CID-W22. The CAT-Peak test attained 95% intelligibility as NU-6 at 27.5 dBHL, and with CID-W22, 92.4% intelligibility at 27 dBHL. The CAT-RMS achieved 90% intelligibility when compared with NU-6, and 87% intelligibility score when compared with CID-W22; all at 24 dBHL.*

callsign test speech intelligibility performance intensity coefficient

1. INTRODUCTION

Speech communication is one of several ways humans interact with the environment. The environment in which we communicate is the limiting factor in our ability to perceive, understand, and recognize the appropriate sound signals transmitted during the interaction process. The environment may be noisy with complex variables affecting the transmission of sound or quiet in some normative sense [1]. Our ability to perform tasks effectively in environments such as the battlefield, airspace management (e.g., pilots and air traffic controllers), hospitals, and manufacturing systems, depends in part on our ability to process speech signals. Effective speech communication requires clear speaking by the talker, a nonrestrictive transmission channel (i.e., medium), and good hearing and speech comprehension by the listener.

Speech intelligibility is a metric for measuring sound or speech signals [2, 3]. Speech intelligibility is critical to every aspect of human communication performance in noisy or quiet environments.

Speech Intelligibility (*SI*) is an index for measuring the minimum absolute threshold of perceiving sound in a given environment. *SI* is quantitatively defined as the percentage of speech units that can be correctly identified by a listener over a given communication system in a given acoustic environment or the degree to which speech can be understood during given conditions [4]. The benefit of high intelligibility is clear. Unintelligible speech is useless and, ironically, low intelligibility speech can be worse; each having the ability to degrade human performance in communication-related tasks. For example, if listeners have to work hard to

Correspondence and requests for offprints should be sent to Celestine A. Ntuen, Institute for Human-Machine Studies, Department of Industrial & Systems Engineering, 419 McNair Hall, 1601 East Market Street, North Carolina A&T State University, Greensboro, NC 27411, USA. E-mail: <ntuen@ncat.edu>.

understand speech, they may become excessively fatigued or they may choose to ignore speech entirely. If speech is easy to misunderstand (as opposed to merely hard to understand), listeners may make more incorrect decisions or incorrect responses in an application. Further, if speech is difficult to understand at any level, listeners may focus so much attention on speech comprehension that they neglect other aspects of their task. In each case, their reactions to speech will be slowed or inappropriate for the intended message, which can have serious consequences in many contexts [5].

Intelligibility tests evaluate the number of words or other speech units that can be correctly identified within a controlled situation. The responses can be objectively scored as a percentage of correct responses [6]. The basic methods of intelligibility testing have been in existence at least since the early 1900s [7, 8, 9]. Whereas *SI* studies and metrics have been applied to many task situations, only recently have interests developed in the military application [10]. This study is the first to attempt to normalize military callsign intelligibility with existing standard tests that have been evaluated to share certain speech characteristics. For this investigation, the Central Institute for the Deaf (CID, Test W-22) and Northwestern University, Auditory Test No. 6 (NU-6) [11] were chosen for the pilot study.

2. SUMMARY OF EXISTING SPEECH INTELLIGIBILITY TESTS

Over the years various metrics and evaluation tools have been developed to measure speech intelligibility. Sometimes, these tools belong to a general class of evaluation metrics known as the articulation index. The articulation index has been widely used as one metric for measuring *SI* [12, 13]. Transmitted speech may be phonemes (the smallest unit of speech), syllables (e.g., consonant-vowel-consonant), words, or sentences. The words and sentences used to test speech intelligibility must be phonetically balanced for a particular language [14]. In such tests, sentence intelligibility is always higher than word intelligibility, and intelligibility for meaningful words is usually higher than intelligibility for meaningless syllables [15]. Some of the standard *SI* tests are summarized here.

2.1. Diagnostic Rhyme Test (DRT)

The Diagnostic Rhyme Test uses a set of isolated words to test for consonant intelligibility in the initial position [14, 16]. The test consists of 96 word pairs that differ by a single acoustic feature in the initial consonant. Word pairs are chosen to evaluate phonetic characteristics. Listeners hear one word at a time, and mark on the answer sheet which one of the two words they think is correct. The DRT does not test any vowels or prosodic features, so it is not suitable for any kind of overall quality evaluation; the test material is quite limited, and the test items do not occur with equal probability; therefore it does not test all possible confusions between consonants. Thus, confusions presented as matrices are difficult to evaluate [17].

2.2. Modified Rhyme Test (MRT)

The Modified Rhyme Test, which is a kind of extension to the DRT, assesses for both initial and final consonant apprehension [14, 16]. The test consists of 50 sets of 6 one-syllable words that make a total set of 300 words. Sets of 6 words are played one at a time, and listeners mark which word they think they hear on a multiple-choice answer sheet. The first half of the words is used for the evaluation of the initial consonants and the second one for the final consonants.

2.3. Standard Segmental Test

The Standard Segmental Test [18, 19] uses lists of consonant, vowel (CV); vowel, consonant (VC); and vowel, consonant, vowel (VCV) nonsense words. All consonants that can occur at the respective positions and three vowels /a/, /i/,

and /u/ are the basic items of the test material. For each stimulus, the missing consonant must be filled on the response sheet, so the vowels are not tested at all. The test material is available with versions in English, German, Swedish, and Dutch.

2.4. Phonetically Balanced Word Lists (PBWLs)

In PBWLs, monosyllabic test words are chosen so that they approximate the relative frequency of phoneme occurrence in each language [14, 16]. The first kind of word list was developed at Harvard University during the Second World War. The relative difficulty of the stimulus items was constrained to items that provided useful information. Several other balanced word lists have been developed since then [14]. For example, the Phonetically Balanced-50 word discrimination test (PB-50) consists of 50 monosyllabic words that approximate the relative frequency of occurrence in English. The PD-100 test was developed to analyze phonetic discrimination and for overall recognition accuracy. The test material includes examples of all possible consonants both in initial and final positions and all vowels in the medial position.

2.5. Helium Speech Intelligibility Testing

The Helium Speech Intelligibility Test (HSIT) was developed by the U.S. Department of Navy to measure and improve the intelligibility of deep-sea divers' voice communications [20]. Helmet gas-flow noise and speech-distorting effects of helium in divers' breathing gas mixture impair the intelligibility of their communications at depth.

2.6. Military Callsign Acquisition Test (CAT)

The Auditory Research Team at the U.S. Army Research Laboratory developed the CAT [4, 10]. It utilizes military callsigns for calling phrases. A single callsign for the CAT consists of a word and a number. The word is a two-syllable military alphabet code and a one-syllable number (e.g., *alpha 1* or *bravo 2*). Listeners are asked to key the callsign they hear through the headphones. The CAT is a speech intelligibility test designed specifically with military applications in mind. It utilizes military callsign alphabets and single number digits to assess speech communication capabilities of various military systems in adverse listening environments. These widely used calling phrases have greater face validity for military applications than speech materials used in any of the existing speech intelligibility tests such as the Modified Rhyme Test (MRT) and the Diagnostic Rhyme Test (DRT). CATs also have greater appeal to soldiers due to their familiarity with test material and task environments. To maintain its ecological validity, it is important to test the CAT in quiet conditions so as to establish a standard and a reference *SI* metric for comparison with results from testing in noisy conditions and with other standard *SI* metrics.

3. METHOD

3.1. Participants

A group of 30 listeners between the ages of 18 and 25 participated in the study. They were recruited from nongovernment civilian and military populations. All listeners had pure-tone hearing thresholds better than or equal to 20 dBHL (decibel hearing level) at audiometric frequencies from 250 through 8,000 Hz (ANSI S3.6-1996) [21] and no history of otologic pathology [22]. An audiometric screening test was performed prior to participation in the study. The screening involved standardized clinical equipment and procedure. The screening facility complied with the ANSI S3.1-1991 [23] requirements for audiometric testing under earphones. After passing the audiometric test, the listeners were asked to sign a consent form and become participants of the study.

3.2. Instrumentation

Instrumentation for the study included (a) a portable IBM PC computer with a CD-ROM drive, (b) a CD-ROM with test material and proprietary CAT software for signal delivery and data collection, (c) a pair of TDH-39 earphones (Telefonics Inc., USA), (d) a Crown A75 power amplifier (Crown, USA) and a step attenuator connected in series between the computer and the earphones, and (e) a KEAMR (Knowles Electronic Manikin for Acoustic Research; Knowles, Inc., USA) simulator and calibration equipment needed to measure sound pressure levels at the ear of the listener.

3.3. CAT

The CAT consists of 126 CAT items. A single item (i.e., a callsign) is a combination of a word selected from a set of 18 two-syllable words comprising the military alphabet (Alpha-Zulu) and a digit selected from a set of 7 one-syllable digits (1 to 8 except 7), for example, Bravo Five. Proprietary CAT software is used [4] to present the test items in randomized order and to record the listeners' responses.

Different sound pressure levels were used in the study. The levels ranged from 5 to 35 dBHL in 5-dBHL steps. A KEMAR manikin was used to determine the audio voltage levels needed to produce respective sound pressure levels at the ear of the listener. Appropriate audio levels were set using the volume control of the Crown power amplifier and an electronic voltmeter connected to the output of the amplifier.

In order to evaluate the effects of equalized peak and equalized average power on the results of CAT administered in a quiet environment, that is, with no noticeable background noise; two versions of CAT were compared: one with equal average power levels of callsign items measured by the Root Mean Square (RMS) power and one with equal maximum power levels of callsign items (Peak). The results from both recordings were compared.

3.4. Procedure

The listener was seated at a station in a sound treated test booth using an IBM PC/586 computer and wearing TDH-39 testing earphones. All the instructions were displayed on the computer screen and the participant was able to use either the computer mouse or the computer keyboard for data input. The listener was asked to listen to a series of CAT items (i.e., military alphabet callsigns and one-syllable numbers from 1 to 8) and to identify them by pressing appropriate keys on the computer keyboard. Figure 1 is an example screen of what the participant viewed while listening to the callsigns.

The test was presented at different sound pressure levels that ranged from 5 to 35 dBHL. The listeners began their first listening period based on the results of their hearing screening. Essentially, the testing began at 5 dB greater than

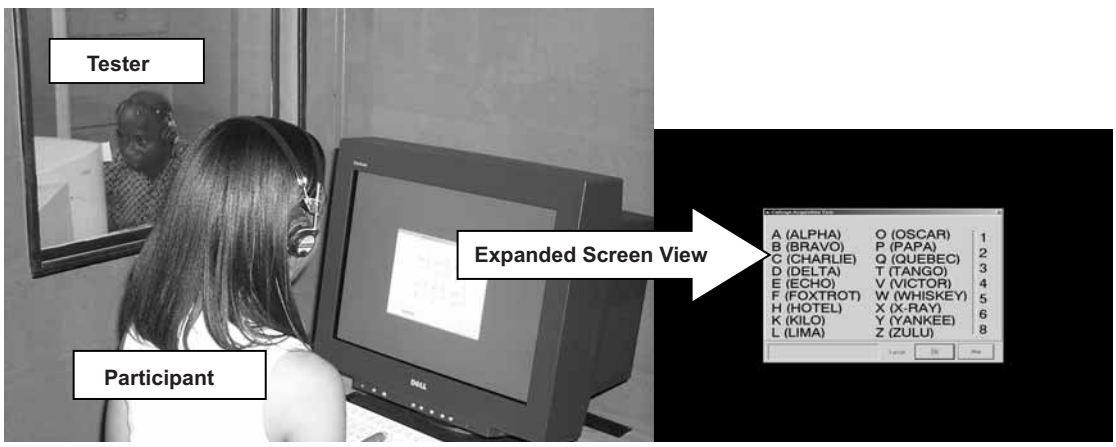

Figure 1. Sample experimental setting for the CAT (Callsign Acquisition Test) speech intelligibility test.

their average pure-tone hearing threshold. Predictably, this was the level at which the participant was expected to score less than 10% correct on either test [10]. The listeners repeated the test with signal level increasing in 5-dB steps until they achieved 95% or better on both RMS and Peak CAT recordings. All the listeners' responses were stored in a file and subsequently imported into an Excel™ spreadsheet for analysis. Each listener participated in a single listening session. The session lasted about 4 hrs and included audiometric screening, instructions, testing, and several 10–15-min long breaks.

4. DATA ANALYSIS

In this analysis, we assume that speech intelligibility (*SI*) is a function of sound pressure levels (*SPL*); that is, in line with the test conditions in which pressure levels were linearly incremented to achieve the 95% test score criterion. First, we model the *SI* function with a nonlinear function of the form [24]

$$SI = I \bullet SPL^{\alpha}. \tag{1}$$

In Equation 1, I is the slope indicating change in *SI* as *SPL* (sound pressure level) changes on the *SI* axis, and α is an unknown constant that is referred to here as word recognition sensitivity or performance intensity coefficient [24]. The parameter α can also be viewed as the spread parameter of the *SI* curve. For analysis, Equation 1 was transformed into a logarithm linear equivalent denoted as

$$\log_{10} SI = I' + \alpha \log_{10} SPL. \tag{2}$$

$\log_{10} I$ is a constant denoted by I'. Equation 2 was analyzed with Statistical Analysis Software (SAS System Setup v5.53.157 Version 8 for Personal Computers) [26]. Tables 1 and 2 are the results obtained for, respectively, Peak and RMS test data.

In Tables 1 and 2, there are two CAT conditions, Peak and RMS; this results in one degree of freedom (*df*). The intercept row gives goodness of fit test statistics for one parameter point intercept of the $\log_{10}$ *SI* scale. These intercept values are 1.098 and –2.09335 in Tables 1 and 2, respectively. The *T* Value and *Pr* > |*t*| are calculated Student *t* statistics and the significant probability of the parameter. The slope (α) is a scalar quantity that measures the change of the speech intelligibility (*SI*) score with respect to changes in *SPL*. The data in the tables show that the regression parameters were significant.

Using these results in Equation 2, Equations 3 and 4 were derived as logarithm *SI* functions for the CATs:

$$\log_{10} SI = -1.098 + 0.137 \bullet \log_{10} SPL$$
$$R^2 = 64.11\% \text{ (Peak)} \tag{3}$$
$$0 \le SPL < \infty,$$

TABLE 1. Parameter Estimates for CAT-Peak Test

Variable	*df*	Parameter Estimates	*SE*	*T* Value	*Pr* > | *t* |
Intercept	1	1.09803	0.02742	40.04	<.0001
Slope (α)	1	0.13709	0.17560	7.81	<.0001

Notes. CAT-Peak—Callsign Acquisition Test with equal maximum power levels of callsign items.

TABLE 2. Parameter Estimates for CAT-RMS Test

Variable	*df*	Parameter Estimates	*SE*	*T* Value	*Pr* > | *t* |
Intercept	1	–2.09335	0.37513	–5.58	<.0001
Slope(α)	1	2.91590	0.30029	9.71	<.0001

Notes. CAT-RMS—Callsign Acquisition Test with equal average power levels of callsign items measured by the Root Mean Square (RMS).

$$\log_{10} SI = -2.093 + 2.916 \bullet \log_{10} SPL$$
$$R^2 = 51.16\% \text{ (RMS)} \qquad (4)$$
$$5 \le SPL < 25.$$

Figure 2 shows the graphical representation of the results obtained from Equations 3 and 4.

Tables 3 and 4 display the descriptive statistics for the two tests at different sound pressure levels.

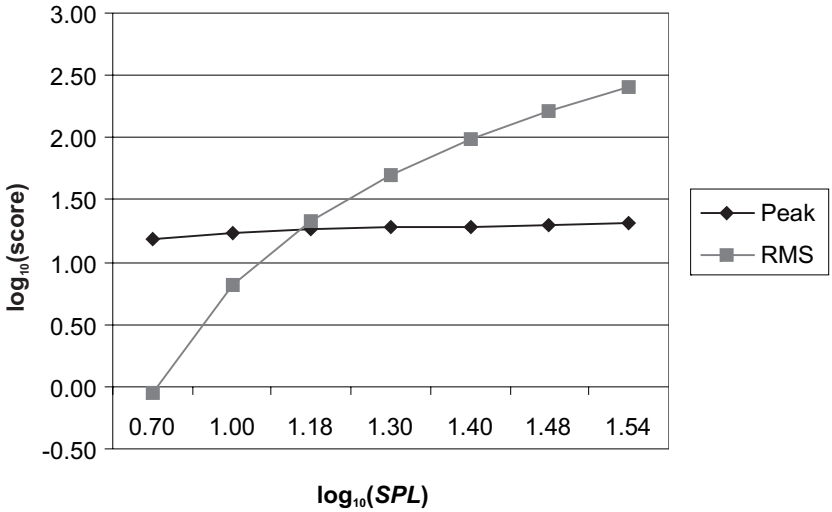

Figure 2. Graph of predicted speech intelligibility for the Callsign Acquisition Test on a log scale. *Notes.* *SPL*—sound pressure level.

TABLE 3. Descriptive Statistics of Scores for CAT-Peak Test

SPL	Average Scores	*SD*	Maximum Score	Minimum Score
10	2.30	0	0	3.97
15	14.59	21.52	0	84.92
20	62.79	34.34	0.79	99.21
25	88.13	18.12	28.57	100.00
30	96.30	4.84	88.10	100.00

Notes. CAT-Peak—Callsign Acquisition Test with equal maximum power levels of callsign items, *SPL*—sound pressure level.

TABLE 4. Descriptive Statistics of Scores for CAT-RMS Test

SPL	Average Scores	*SD*	Maximum Score	Minimum Score
5	3.17	0	3.17	3.17
10	4.85	7.47	26.98	0
15	44.01	32.35	96.83	0
20	79.61	30.53	100.00	0.79
25	94.89	8.31	100.00	70.63
30	94.84	8.04	100.00	79.37
35	100.00	0	100.00	100.00

Notes. CAT-RMS—Callsign Acquisition Test with equal average power levels of callsign items measured by the Root Mean Square (RMS), *SPL*—sound pressure level.

5. STANDARDIZING THE CAT USING NU-6 AND CID W-22 METRICS

Wilson and Oyler [11] completed a study comparing the Central Institute for the Deaf (CID W-22) and Northwestern University Auditory Test No. 6 (NU-6). In this study, the material for both tests was presented in quiet conditions (no background noise) at 0–30 dBHL to 24 different listeners with normal hearing. The results from their study were used in this analysis as a basis for comparison. The mean percent correct recognition data and standard deviation for these studies are listed in Table 5.

For comparison, the mean scores for all four tests were used. Figure 3 shows all four *SI* curves for two versions of the CAT and CIDW-22 and NU-6 tests, respectively. From Figure 3, it can be concluded, based on the results obtained earlier by Wilson and Oyler [11], that *SI* performance on

TABLE 5. The Average Percent Correct Recognition and Standard Deviation for NU-6 and CID W-22

	NU-6		CID W-22	
dBHL	***M***	***SD***	***M***	***SD***
0	3.7	4.6	0.4	1.0
5	13.6	10.6	6.5	5.2
10	35.1	18.6	22.7	14.3
15	58.2	18.2	47.5	20.5
20	78.0	11.5	72.6	14.1
25	87.7	6.4	84.8	6.9
30	93.2	3.5	90.5	4.5

Notes. dBHL— decibel hearing level, CID W-22—Central Institute for the Deaf, NU-6—Northwestern University Auditory Test No. 6.

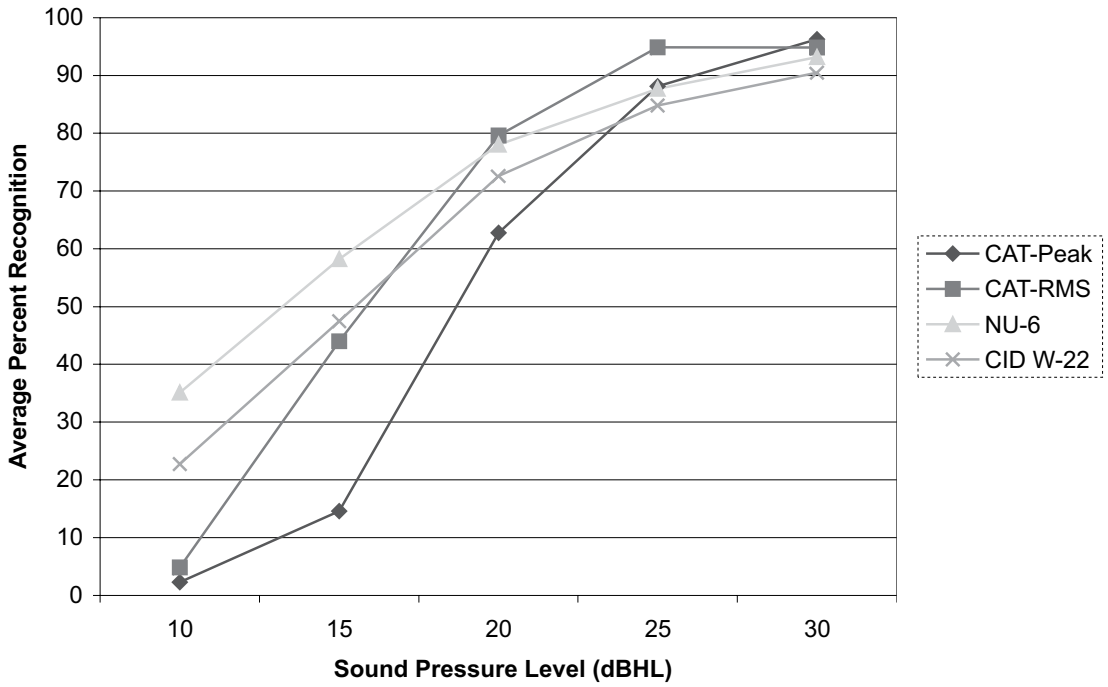

Figure 3. Speech Intelligibility curves for CAT-Peak and CAT-RMS, CIDW-22 and NU-6. *Notes*. CAT-Peak—Callsign Acquisition Test with equal maximum power levels of callsign items, CAT-RMS—Callsign Acquisition Test with equal average power levels of callsign items measured by the Root Mean Square (RMS), dBHL— decibel hearing level, CID W-22—Central Institute for the Deaf, NU-6—Northwestern University Auditory Test No. 6.

both the Peak and RMS versions of the CAT were slightly more difficult, but the four tests were comparably similar. The CAT could possibly replace either CIDW-22 or NU-6 tests in clinical or military environments. These results were validated by the correlation coefficients: NU-6 and CAT-RMS = 95.1%, NU-6 and CAT-Peak = 96.78%, CID W-22 and CAT-RMS = 99.47%, and CID W-22 and CAT-Peak = 97.57%.

To compare the CAT against NU-6 and CID-W22 tests, mean data for NU-6 and CID W-22 performance scores were regressed against the two different CATs. The results obtained are as follows:

(a) Comparing CAT-Peak test and NU-6:

$$\begin{gathered} SI_{(\text{CAT-Peak})} = 1.73 \bullet SI_{\text{NU-6}} - 68.96 \\ R^2 = 93.66\% \\ 0 \le SI_{(\text{CAT-Peak})} \le 100 \\ 39.86 \le SI_{\text{NU-6}} \le 97.66. \end{gathered} \tag{5}$$

(b) Comparing CAT-Peak test and CID-W22:

$$\begin{gathered} SI_{(\text{CAT-Peak})} = 1.47 \bullet SI_{\text{CID-W22}} - 40.823 \\ R^2 = 95.21\% \\ 0 \le SI_{(\text{CAT-Peak})} \le 100 \\ 27.77 \le SI_{\text{CID-W22}} \le 95.79. \end{gathered} \tag{6}$$

(c) Comparing CAT-RMS test with NU-6:

$$\begin{gathered} SI_{(\text{CAT-RMS})} = 1.62 \bullet SI_{\text{NU-6}} - 50.71 \\ R^2 = 99.02\% \\ 0 \le SI_{(\text{CAT-RMS})} \le 100 \\ 31.3 \le SI_{\text{NU-6}} \le 93. \end{gathered} \tag{7}$$

(d) Comparing CAT-RMS with CID-W22:

$$\begin{gathered} SI_{(\text{CAT-RMS})} = 1.37 \bullet SI_{\text{CID-W22}} - 23.54 \\ R^2 = 98.95\% \\ 0 \le SI_{(\text{CAT-RMS})} \le 100 \\ 17.18 \le SI_{\text{CID-W22}} \le 90.2. \end{gathered} \tag{8}$$

These relationships are graphically illustrated in bivariate plots of CAT scores versus NU-6 and CID-W22 test scores as shown in Figures 4 and 5.

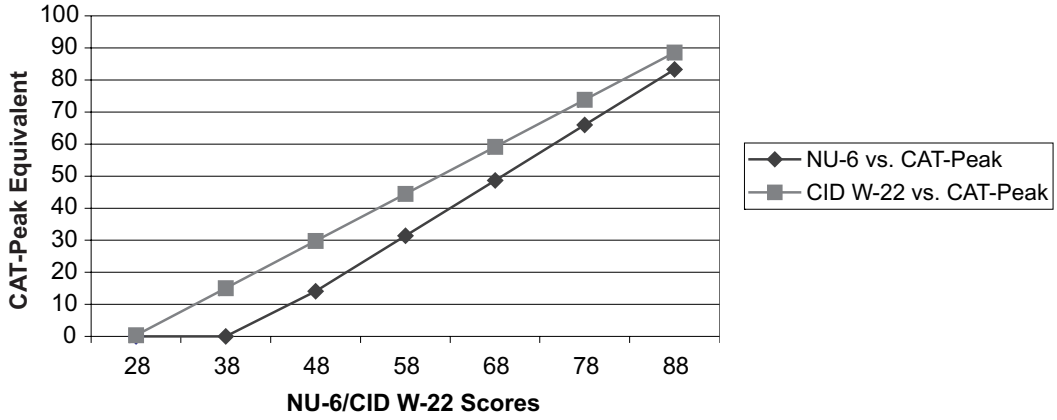

Figure 4. Bivariate plots of CAT-Peak Speech Intelligibility (*SI*) scores against the NU-6 and CIDW-22 *SI* scores. *Notes*. CAT-Peak—Callsign Acquisition Test with equal maximum power levels of callsign items, CID W-22—Central Institute for the Deaf, NU-6—Northwestern University Auditory Test No. 6.

TABLE 6. Equivalent Standard Test Scores on Callsign Acquisition Tests (CATs)

	CAT-Peak Test		CAT-RMS Test	
	Minimum *SPL*	Maximum *SPL*	Minimum *SPL*	Maximum *SPL*
Test Scores	(10.57)	(28.42)	(5.57)	(26.5)
CAT	0%	100%	0%	100%
NU-6	39.86%	97.66%	31.30%	93.00%
CID-W22	27.77%	95.79%	17.18%	90.20%

Notes. CAT-Peak—Callsign Acquisition Test with equal maximum power levels of callsign items, CAT-RMS—Callsign Acquisition Test with equal average power levels of callsign items measured by the Root Mean Square (RMS), *SPL*—sound pressure level, CID W-22—Central Institute for the Deaf, NU-6—Northwestern University Auditory Test No. 6.

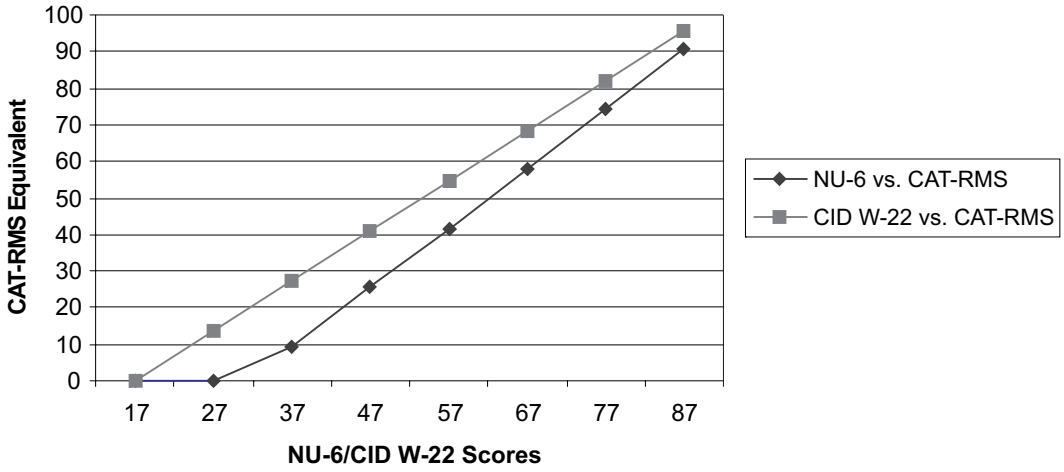

Figure 5. Bivariate plots of CAT-RMS Speech Intelligibility (*SI*) scores against the NU-6 and CIDW-22 *SI* scores. *Notes*. CAT-RMS—Callsign Acquisition Test with equal average power levels of callsign items measured by the Root Mean Square (RMS), CID W-22—Central Institute for the Deaf, NU-6—Northwestern University Auditory Test No. 6.

As seen in Figures 4 and 5, NU-6 leads CID W-22 tests in both Peak and RMS tests by 10% *SI* scores. Thus, it can be said that the NU-6 test battery is 10% easier than CIDW-22. The equivalent standard scores represent the validity of using the CAT as a comparative speech intelligibility assessment tool in place of the NU-6 and CID-W22 tests. In this case, for a quiet environment 0% *SI* score on the CAT-Peak test is equivalent to a 39.86% score on NU-6. A summary of the comparison is presented in Table 6.

6. CONCLUSION AND DISCUSSIONS

The data obtained for the CAT were compared with NU-6 and CID-W22 tests. The purpose of the comparison was to derive normalizing parameters for using CATs. The following results were obtained:

1. For the CAT-Peak test under a minimum threshold *SPL* value of 10.57 dBHL, 0% intelligibility score was obtained, whereas 39.86% was obtained for NU-6 and 27.77% for CID-W22;
2. For the CAT-Peak test under a maximum *SPL* value of 28.42, 100% intelligibility score was obtained, whereas 97.66% was obtained for NU-6 and 95.79% for CID-W22;
3. For the CAT-RMS test with a minimum *SPL* threshold of 5.57 dBHL, we obtained a 0% intelligibility score, whereas 31.3% was obtained for NU-6 and 17.18% for CID-W22; and
4. For the CAT-RMS test under maximum *SPL* value of 26.5 dBHL, a 100% intelligibility score was obtained, whereas 93% was obtained for NU-6 and 90.2% for CID-W22.

From these results, it can be inferred that at a minimum *SPL* of 10.57 dBHL, CATs are more difficult when compared to NU-6 and CID-W22. Similarly, it appears that CID-W22 is more difficult than NU-6. At the maximum *SPL* values, CATs reveal more intelligibility than both NU-6 and CID-W22 (i.e., approximately 7% better). Comparing the CAT with NU-6 and CID-W22 at an intelligibility score of 95% revealed the following: a CAT-Peak test will attain 95% like NU-6 at 27.5 dBHL; and with CID-W22, it will attain 92.4% at 27 dBHL. The CAT-RMS will achieve a 90% intelligibility score when compared with NU-6, and an 87% intelligibility score when compared with CID-W22; all at 24 dBHL. These values give standard CAT speech discrimination thresholds in a quiet environment.

There are at least four ergonomic and safety consequences of this research:

1. The analytical relationships can be used to establish validity and robustness of NU-6 and CID-W22 in intelligibility test applications.
2. The performance intensity (*PI*) coefficient is a useful parameter in establishing the just-noticeable sound perception threshold. The *PI* coefficient can be used to establish initial speech perception threshold. The result can be useful in reducing potential workload associated with listening stress [26].
3. The standard metric can be used to compare different speech intelligibility at different listening conditions.
4. With more research, the standard metric can be used to establish correlation of speech intelligibility and speaker audibility. This can be done through explicit inference on the used 95% intelligibility significant score.

REFERENCES

1. Frank H, Karlovich RS. Effect of contralateral noise on speech detection and speech reception thresholds. Audiology 1975;14:34–43.
2. Miller GA, Neovius L, Raghavendra P. An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 1955;27(2):338–52.
3. Owens E, Schbert ED. The development of constant items for speech discrimination testing. J Speech Hear Res 1968;11;656–67.
4. Letowski T, Karsh R, Vause N, Shilling R, Ballas J, Brungart D, et al. Human factors military lexicon: Auditory displays [unpublished technical report]. Aberdeen Proving Grounds, MD, USA: U.S. Army Research Laboratory, Human Research Engineering Directorate; 2001.
5. Gardner-Bonneau D, editor. Human factors and voice interactive systems. Hingham, MA, USA: Kluwer Academic; 1999.
6. Sydral A, Bennett R, Greenspan S. Applied speech technology. Boca Raton, FL, USA: CRC Press LLC; 1994.
7. Fletcher H. Speech and hearing in communication. Princeton, NJ, USA: Van Nostrand Reinhold; 1953.
8. Fletcher H, Steinberg JC. Articulation testing methods. Bell Systems Technical Journal 1929;7:806–54.
9. Speaks C, Jerger J. Performance intensity characteristics of synthetic sentences. J Speech Hear Res 1966;9:305–12.
10. Letowski T. Performance intensity function for the Callsign Acquisition Test (CAT) research protocol [unpublished technical report]. Aberdeen Proving Grounds, MD, USA: U.S. Army Research Laboratory, Human Research Engineering Directorate; 2001.
11. Wilson R, Oyler A. Psychometric functions for the CID W-22 and NU auditory Test No. 6. Materials spoken by the same speaker. Ear Hear 1997;18(5):430–4.
12. Kamm C, Dirks D, Bell T. Speech recognition and the articulation index for normal and hearing-impaired listeners. J Acoust Soc Am 1985;77:281–8.
13. Summers B. Speech synthesis. J Acoust Soc Am 1988;84:917–28.
14. Goldstein M. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. Speech Communication 1995;16: 225–44.
15. Bailey RW. Human performance engineering: designing high quality professional user interfaces for computer products, applications and systems. 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall; 1996.
16. Logan J, Greene B, Pisoni D. Segmental intelligibility of synthetic speech produced by rule. J Acoust Soc Am 1989;86(2): 566–81.

17. Carlson R, Granström B, Nord L. Evaluation and development of the KTH text-to-speech system on the segmental level. In: Proceedings of International Conference of Acoustics, Speech, and Signal Processing. Woodbury, NY, USA: Acoustical Society of America; 1990. p. 317–20.

18. Jekosch U. Speech quality assessment and evaluation. In: Jiri S, Vratislav D, editors. Proceedings of the Third European Conference on Speech Communication and Technology "Eurospeech '93". Berlin, Germany: European Speech Communication Association; 1993. p. 1387–94.

19. Pols L. Multilingual synthesis evaluation methods. In: Stanton R, editor. Proceedings of International Conference on Spoken Language Processing, ICSLP '92. Edmonton, Alta., Canada: Quality Color Press; 1992. p. 181–4.

20. Hamill BW. Helium speech intelligibility testing in a noisy saturation diving environment. In: 3rd Symposium on Research & Development, Johns Hopkins University, Applied Physics Laboratory [abstract]; 1995. Retrieved May 5, 2004, from http://www.jhuapl.edu/symposium/3rd_RandD/Helium.htm

21. American National Standards Institute (ANSI). Specifications for audiometers (Standard No. ANSI S3.6-1996). New York, NY, USA: ANSI; 1996.

22. Mendel LL, Danhauer JL. Audiological evaluation and management of speech perception assessment. San Diego, CA, USA: Singular; 1997.

23. American National Standards Institute (ANSI). Maximum permissible ambient noise levels for audiometric test rooms (Standard No. ANSI S3.1-1991). New York, NY, USA: ANSI; 1991.

24. Blue M, Ntuen CA. Performance intensity (PI) function of callsign acquisition test (CAT). In: Bidanda A, editor. Proceedings of the Institute of Industrial Engineering Research Conference [CD-ROM]. Atlanta, GA, USA: Institute of Industrial Engineer Management Press; 2003. p. 84–99.

25. Beattie RC, Svihovec V, Edgerton BJ. Relative intelligibility of the CID spondees as presented via monitored live voice. Journal of Speech Hearing Disorder 1975;40:84–91.

26. Cody RP, Smith JK. Applied Statistics and the SAS programming language. 4th ed. Upper Saddle River, New Jersey: Prentice Hall; 1997.

27. Cooper FS, Delattre PC, Liberman AM, Borst JM, Gerstman LJ. Some experiments on the perception of synthetic speech sounds. J Acoust Soc Am 1952;24:597–606.