Evaluation of a Protocol for Integrated Speech Audiometry

Jerry Punch and Brad Rakerda

Purpose: This project was aimed at evaluating the reliability, validity, and clinical utility of a protocol for integrated measurements of the most comfortable level (MCL) and uncomfortable level (UCL) for speech, in combination with the speech recognition threshold (SRT). We also evaluated the validity of using spondee words when measuring speech MCL and UCL.

Method: In a randomized block design, equal numbers of women and men with normal hearing, aged 18–29 years, were assigned to each of 3 experimental stimulus conditions: spondee singlets, spondee triplets, or connected discourse (n = 12 per group). Following measurement of the SRT, a modified method of limits was employed to establish, on a 7-point loudness rating scale, an ascending MCL, a descending MCL, and an ascending UCL. A single instructional set covered all loudness measurements. Test times were tracked electronically to assess clinical efficiency. All test conditions were repeated during each of 2 separate test sessions.

Results: Mean SRTs, MCLs, and UCLs across the 3 different experimental groups were found not to differ statistically or clinically (mean differences < 5 dB). Intrasession and intersession reliability for the various measures were excellent, and testing of all listeners was completed in a timely manner. In a follow-up experiment with adults with normal hearing who were only a decade older than participants in our main experiment, the older group was found to have significantly higher MCLs and UCLs.

Conclusions: Spondee words can be used routinely to obtain reliable, valid, and clinically efficient measures of MCLs and UCLs for speech, in a protocol combined with the SRT. Spondees, presented singly, yielded the greatest level of efficiency overall. Results support a recommendation to obtain an ascending measurement of MCL prior to a descending measurement and to establish the MCL by averaging the 2 values.

Speech audiometry typically consists of determining the speech recognition threshold (SRT), the word recognition score (WRS), the most comfortable loudness level (MCL) for speech, and the uncomfortable loudness level (UCL) for speech. The WRS is established in quiet and, preferably, also in noise. Textbook descriptions of these measures generally promote the view that audiologists who wish to perform best practices with their clinical patients should include these tests as critical components of routine hearing evaluations and hearing aid evaluations (e.g., Gelfand, 2009; Martin & Clark, 2015).

The SRT serves as a single numerical value indicating the lowest level at which speech is recognized. Audiologists generally regard it, minimally, as a basis for characterizing the overall degree of hearing loss and as a means of confirming the reliability of pure-tone audiometric thresholds in the primary speech frequencies of 500–2000 Hz. It can also be used as a reference for determining the level at which suprathreshold speech recognition testing is performed; as a means of assessing the need for, as well as benefit from, hearing aids or other rehabilitative measures; and as a simple diagnostic tool for identifying nonorganic hearing loss.

The MCL, in addition to establishing the level at which a patient prefers to listen to speech material, has traditionally been considered by many audiologists to be the optimal level at which to obtain the WRS (see Punch, Joseph, & Rakerd, 2004, for a review). Historically, another use of the MCL is as an estimate of the gain needed in a hearing aid to produce comfortably loud speech when speech is presented at 40 dB HL (Carhart, 1946a, 1946b; Shapiro, 1976) or 50 dB HL (Ventry & Johnson, 1978). Using Carhart’s assumption that speech is most comfortable at 40 dB HL, the applicable formula is

\[
\text{Gain} = \text{MCL}_{-40\text{dBHL}} - 40\text{dBHL}.
\] (1)

More recently, the MCL has been used to establish what is known as the acceptable noise level (ANL). The
ANL is a test of noise tolerance, the amount of background noise that a listener is willing to accept while listening to running speech (Nabelek, Tucker, & Letowski, 1991). It is essentially the signal-to-noise ratio that represents the difference between the MCL and the background noise level that is varied during the test. The ANL has been shown to be predictive of the successful use of, as well as potential benefit from, hearing aids (Nabelek, Freyaldenhoven, Tampas, Burchfield, & Muenchen, 2006).

The UCL, the level at which listeners perceive speech to be uncomfortably loud, is used to estimate the output sound pressure level of a hearing aid, as it is widely accepted that an aid’s output should not exceed a listener’s UCL. Finally, the difference between the UCL and the SRT represents the dynamic range of speech for individual listeners.

Despite the fact that many audiologists continue to include MCL and UCL measurements routinely in their speech audiometry battery, the materials and procedures for administering tests of the MCL and UCL are not well standardized. A systematic literature review (Punch, Joseph, et al., 2004) found that test methods and instructions strongly influence both the MCL and the UCL, whereas stimulus conditions appear to affect them less substantially. The literature also suggests that the MCL is best considered as a range of values rather than a fixed level and that the MCL is typically higher when assessed with a descending approach, as opposed to an ascending approach. Several studies (Kavanagh, 1960; Schaanenman, 1965; Ventry & Johnson, 1978; Wall & Gans, 1984; Woods, Ventry, & Gatling, 1973) have reported the MCLs obtained by descending methods to be as much as 15–20 dB higher than those obtained by ascending methods.

The MCL has also been shown to be higher when preceded by measurement of the UCL. Using spondaic words as stimuli, Punch, Rakerd, and Joseph (2004) demonstrated experimentally that the mean MCL in listeners with normal hearing was about 9 dB lower when it preceded the UCL measurement than when it followed the UCL measurement. They also found that test order and temporal spacing during tests of the MCL and UCL exert a powerful influence on the MCL, but not on the UCL. Specifically, they found that the MCL was almost 7 dB higher on retest, following reinstruction and exposure to higher levels during a test of the UCL. If the UCL measurement immediately preceded the MCL measurement, the MCL was 11–13 dB higher than when UCL testing followed MCL testing.

Our literature review (Punch, Joseph, et al., 2004) and experimental study (Punch, Rakerd, et al., 2004) led us to make several recommendations regarding measurements of the MCL and UCL for speech. The first was that a best numerical estimate of the MCL should be obtained by averaging the results of both ascending and descending runs. The second was that the ascending MCL measurement (MCL-A) should be made before the descending MCL measurement (MCL-D). Additional recommendations were that the MCL measurement should generally precede the UCL measurement and that instructions should clearly indicate the basis for a listener’s loudness judgments. Here, we implemented all of these recommendations in a computer-controlled test protocol that integrated measurements of the speech MCL and UCL with a measurement of the SRT. The primary purpose of this study was to evaluate the clinical effectiveness of this integrated protocol. Accordingly, we report below on its reliability, concurrent validity, and timeliness.

The study also had a second purpose, which was to determine whether the spondees routinely used to find a listener’s SRT could also be used clinically to establish the MCL and UCL, within the framework of the above-described integrated protocol. Spondaic words, usually presented singly, have often been used for MCL and UCL measurements in the research literature (e.g., Beattie, Edgerton, & Gager, 1979; Edgerton, Beattie, & Wides, 1980; Punch, Joseph, et al., 2004; Sammeth, Birman, & Hecox, 1989). A primary rationale for believing that spondees might replace connected speech in clinical measurements of the MCL and UCL is that various metrics used to compare the two types of speech stimuli have demonstrated more commonalities than differences. Speaks (1967), for example, found similar performance-intensity functions and 50% correct responses for sentences and spondaic words in quiet. He also found a difference of only about 2 dB in the range spanning 20%-80% intelligibility for synthetic sentences and spondaic words in quiet. Dirks, Wilson, and Bower (1969) found the response range from 20% to 80% to encompass an average of about 5 dB for sentences and 6 dB for spondaic words during continuous noise conditions. Under a variety of listening conditions involving continuous and interrupted maskers, they found a high degree of similarity in the results for spondees and sentences. These results confirm that the performance-intensity functions for the two types of stimuli are highly similar to one another.

In this study, we employed an automated protocol that closely followed commonly used clinical measurement methods and used it to compare the MCL and UCL results obtained with (a) spondees presented singly, (b) spondees presented in a three-word series per trial, and (c) connected discourse. To establish the clinical viability of the protocol, we conducted both intrasession and intersession assessments of the method’s reliability and tracked the time needed to complete all test procedures.

**Method**

**Test Room**

Participants were tested individually in a double-walled audiometric sound room. All SRT, MCL, and UCL testing
was computer controlled. During testing, participants sat facing a standard computer monitor and used a computer mouse to select and enter all of their responses.

**Test Sequence**

To establish both intrasession and intersession reliability of all measurements, we conducted test and retest sessions during both an initial day of testing (Day 1) and a follow-up session approximately 2 weeks later (Day 2). The order of testing on Day 1 was (a) familiarization with a set of 36 spondee words (American Speech-Language-Hearing Association [ASHA], 1988) to control for any effects of prior knowledge of the vocabulary (Tillman & Jerger, 1959), (b) SRT testing, (c) combined testing of the MCL and UCL (with the MCL preceding UCL testing), (d) an approximately 10-min break, (e) SRT retesting, and (f) combined MCL/UCL retesting. The same test–retest sequence was followed on Day 2, with the exception that familiarization was not repeated. Michigan State University’s Institutional Review Board approved all aspects of the study.

**Participants**

Thirty-six participants (18 women and 18 men), aged 18–29 years, were recruited from the campus of Michigan State University and from the greater Lansing, Michigan, community. All were accomplished speakers of English, and all exhibited normal hearing by passing an audiometric screening at 20 dB HL at 500, 1000, 2000, and 4000 Hz bilaterally. All experimental testing was conducted in each participant’s preferred ear.

**Experimental Groups**

A randomized block design was used, in which six women and six men were randomly assigned to one of three experimental groups. The three groups—referred to here as **singlets, triplets, and discourse**—were presented with different sets of stimuli during the measurement of the MCL and UCL. The singlets group was presented with the same spondees used for the measurement of the SRT, with those spondees presented singly (one randomly selected spondee per trial), as done in a standard SRT test. The triplets group was also presented with spondees, but with a series of three spondees presented on each trial (three randomly selected spondees, presented at the rate of one per second). We used spondee triplets here because they more closely mimic the length of connected discourse than do single spondees, and we wished to compare their use with both spondee singlets and connected discourse as measures of the MCL and UCL. Wall and Gans (1984) also used spondee triplets and found that they are a viable stimulus in measuring the MCL for speech. In this study, we used triplets to measure both the MCL and UCL. Finally, our discourse group was presented with a series of continuously spoken phrases and sentences, described below.

**Stimuli**

**Spondees**

To establish the SRT for all three experimental groups and to establish the MCL and UCL for the singlets and triplets groups, we used spondee words produced by a single male talker with a general American accent. The talker produced 36 spondees (ASHA, 1988), which were digitally recorded (24 bits per sample, sample rate = 44.1 kHz) and then adjusted in level so that they all had the same root-mean-square power. A 1000-Hz tone with that same power was used as a calibration signal.

**Discourse**

For the discourse group, the stimuli used to establish the MCL and UCL were 31 different phrases and sentences, all excerpted from *The Rainbow Passage* (Fairbanks, 1927). These discourse stimuli were spoken by the same male talker who produced the spondees. Each phrase or sentence was digitally recorded and then adjusted in level so that its root-mean-square power matched that of the spondees.

**Instructions**

**Familiarization and SRT**

For familiarization and SRT measurement, participants were given written instructions recommended by Gelfand (2009, p. 242), adapted to accommodate an automated procedure in which participants indicated their recognition of spondees by the use of a computer mouse and the response interface pictured in Figure 1. Appendices A and B provide the instructions for familiarization and SRT testing, respectively.

**MCL and UCL**

Written instructions were combined for the MCL and UCL measurements. Participants were told that the overall purpose of the MCL/UCL test was to obtain their judgments of how comfortable words were when heard at different levels or volume settings. Depending on the experimental group, participants were informed that they would hear a series of two-syllable words, presented one at a time or three at a time, or that they would hear continuous speech. They were also told that the speech volume would become progressively louder and softer and that they were to rate the loudness of each speech sample and report the rating by clicking on a number from 1 to 7 using the response interface shown in Figure 2. The interface, mirroring that used in the Contour Test of Loudness Perception (Cox, 1995; Cox, Alexander, Taylor, & Gray, 1997), was visible throughout the test, and it defined the seven loudness rating steps as follows: 1 = very soft, 2 = soft, 3 = comfortable but slightly soft, 4 = comfortable, 5 = comfortable but slightly loud, 6 = loud but OK, and 7 = uncomfortably loud. A final instruction to participants was to pretend that they were listening to the voice of a radio broadcaster and to keep in mind that a rating of 7 indicated that the broadcast was louder than they...
would ever choose when listening to the radio. Appendix C contains the specific instructions given to the singlets group; similar wording was used for the other two groups, with minor wording changes that were specific to the stimulus condition administered.

**Psychophysical Method**

**SRT**

SRT testing was included in the present protocol for three reasons. First, the SRT is a routine component of speech audiometry in audiologic evaluations. Second, the spondees used to establish the SRT here were also used in two of the three stimulus conditions of the experiment (singlets and triplets). Third, the SRT found on a run determined the starting level for subsequent testing, as MCL/UCL testing was initiated at the SRT plus 20 dB.

The SRT was established by first familiarizing each participant with the 36-item corpus of spondee words. During familiarization, the spondees were presented at a uniform level of 50 dB HL. Computer programming allowed tracking of any words missed, and only words correctly recognized during the familiarization process were used as stimuli during SRT testing. Following familiarization, the 5-dB ASHA (1988) method was used to establish the SRT.²

²Across all 36 participants, who had normal hearing, there was only one instance of a missed word during familiarization.

²The ASHA 5-dB method was chosen arbitrarily, as neither the 2- nor 5-dB method has been shown to offer a time savings advantage over the other and both have been shown to yield comparable results (Chaiklin & Ventry, 1964).

**MCL and UCL**

On each run, a modified method of limits search yielded an MCL-A, an MCL-D, and an ascending measurement of the UCL.

**MCL-A.** This was established by initiating, in the test ear, an ascending method of limits search at a level 20 dB above the SRT. We presented the group-dependent stimulus (singlets, triplets, or discourse) in 5-dB ascending steps, storing the participants’ ratings (1–7) for each stimulus. At the first instance of a response of *comfortable but slightly loud* (5) or higher, the level was reduced by 10 dB, and another stimulus presented at that level, in a sequence that continued to decrease in 10-dB steps until a response of *comfortable but slightly soft* (3) or below was obtained. We then increased the volume in 5-dB steps, and this cycle (5-dB increments and 10-dB decrements) was repeated until two ratings of *comfortable* (4) were obtained at the same level during three consecutive ascending presentations. We stored this response as the MCL-A. The word prompts *Pause* and *Resume* (after a 5-s pause) were shown on the computer monitor to indicate to participants that a new test (MCL-D) was about to begin.

**MCL-D.** We next established MCL-D by initiating a descending method of limits search at a level 20 dB above the MCL-A. The group-dependent stimulus was presented in 5-dB descending steps until a rating of *comfortable but slightly soft* (3) or lower was achieved, at which point the level was increased in 10-dB steps until a response of *comfortable but slightly loud* (5) or higher was obtained. This cycle (5-dB decrements and 10-dB increments within the relevant range) was repeated until two ratings of *comfortable* (4) were obtained at the same level within three consecutive descending presentations. That level was stored as the MCL-D. Again, the words *Pause* and *Resume* were shown on the computer monitor to indicate to participants that a new test (the UCL search) was about to begin.
**Results and Discussion**

**Interpretation of Significance of Differences**

A concern in evaluating the outcomes of this study was that of determining what constitutes a clinically significant difference between specified measures. A variety of indices exists, of course, to specify significance of differences, including tests of statistical significance, tests of effect size, and clinical norms. We will refer to all of these indices in this section, but a knowledge of the research evidence has led us to focus particularly on clinical norms that often treat test–retest and other differences of greater than 5 dB as clinically significant in such measures as pure-tone and speech audiometric thresholds. The fact that audiometer attenuators have traditionally been demarcated and calibrated to the nearest 5 dB appears to be a de facto indication that clinical audiologists regard any audiometric difference of less than 5 dB as clinically insignificant. For example, audiologists have traditionally performed a reliability check of pure-tone thresholds at 1000 Hz to confirm that pure-tone thresholds, in general, are reliable within 5 dB. Gelfand (2009) states:

A given clinical threshold measurement is generally considered to be reliable within ± 5 dB (p. 142). Pure-tone thresholds are usually repeatable within a range of ± 5 dB, and test-retest reliability is certainly expected to be within ± 10 dB. (p. 411)

In the following subsections, we will consider both the statistical and clinical significance of any differences in light of the above-described clinical norms.

**SRT**

The mean SRT of participants in this study was 1.7 dB HL. Low SRTs were expected, given that participants were young adults who had passed a hearing screening at 20 dB HL. Mean SRTs and standard deviations (in parentheses) for the different stimulus groups were as follows: singlets, 1.3 dB (3.1 dB); triplets, 1.9 dB (3.2 dB); and discourse, 1.9 dB (2.4 dB). The reliability of the SRT measurements was high, both within and across days. On Day 1, there was only one instance (out of 36) in which a participant’s test and retest SRT differed by more than 5 dB, and on average (over the 36 participants), the test–retest difference

---

**MCL range and MCL.** We refer here to the difference between MCL-A and MCL-D as the *MCL range*, and we calculated it on each run as MCL-D minus MCL-A. The participants’ MCLs were calculated on each run by averaging the MCL-A and MCL-D.

**UCL.** The UCL was established by an ascending method of limits search that began at the MCL for the run (rounded down to the nearest 5 dB). The group-dependent stimulus was presented in 5-dB ascending steps, and the participants’ ratings (1–7) for each stimulus was stored. At the first instance of a response of “uncomfortably loud” (7), the level was reduced by 10 dB, with a stimulus presented at that level, and this procedure was continued by reducing the level in 10-dB steps until a response of “comfortable but slightly loud” (5) or below was obtained. The volume was then increased in 5-dB steps, and this cycle (5-dB increments and 10-dB decrements) was repeated until two ratings of “loud but OK” (6) were obtained at the same level during three consecutive ascending presentations. The level at which this rating was obtained was defined as the UCL for the run.

As in our previous work, we define the UCL as the level at which speech is “loud but OK,” not the level at which speech becomes truly uncomfortable. This is essentially the level described by Cox (1988) as the upper limit of comfortable loudness and by Mueller (2011) as the upper level of comfort. Our primary reason for not establishing a truly uncomfortable level was the wish to minimize exposure to levels that could result in repeated discomfort or pain. We view our definition of the UCL and its measurement as consistent with that expressed by Mueller and Bentler (2005), who noted that a popular clinical procedure is to measure aided loudness discomfort, or loudness discomfort level, by verifying that high input levels are “loud, but okay.” They state:

Most clinics use a protocol similar to the IHAFF verification approach (Cox, 1995; Mueller, 1999). In this case, the goal of the test is not to determine LDL per se but, rather, deliver high-level inputs to the patient (e.g., speech or environmental noises at 85 dB SPL), which on average should be judged as “Loud, But Okay” (just below the LDL). If these high-level inputs are judged to be either too loud or too soft, the output of the hearing aid is adjusted accordingly. (p. 465)
was just +0.1 dB. On Day 2, there was again just one instance of a test-retest difference greater than 5 dB, and the average difference was −0.7 dB. Finally, for every participant, the SRT measured on Day 2 was within 5 dB of the SRT measured on Day 1. The average difference between the Day 2 and Day 1 SRTs was +0.6 dB.

MCL-A, MCL-D, and MCL Range

At least two previous studies have found a gender difference when rating loudness, with women having significantly lower intensity levels than men for equivalent loudness ratings (Kiessling, Steffens, & Wagner, 1993; Nielsen, 1995). In this study, we found no such difference between the women’s and men’s results for either the MCL-A, \( t(34) = 1.209, p = .235 \), or the MCL-D, \( t(34) = 0.428, p = .672 \), or for any of the other comparisons to be reported below (\( p > .05 \) in all cases). We pooled the women’s and men’s results, therefore, when assessing all effects of stimulus type.

Figure 3 shows the MCL-A and MCL-D results of the experiment, plotted separately for the three stimulus groups. The individual participants’ test results were first averaged over four runs (two runs on Day 1 and two runs on Day 2) to get the best estimates of their MCL-A and MCL-D. Those level data were then averaged over the participants in each group to obtain the means and standard deviations reported in the figure. Analysis of variance revealed no significant difference among the three different stimulus group means for either the MCL-A, \( F(2, 33) = 1.016, p = .373 \), or the MCL-D, \( F(2, 33) = 0.454, p = .639 \). An effect size analysis found that stimulus type accounted for only about 3% of the variance for the MCL-D (\( \eta^2 = .027 \)) and about 6% for the MCL-A (\( \eta^2 = .058 \)).\(^4\) The largest mean difference between either of the spondees and the discourse group was just 2.7 dB for the MCL-A and 1.9 dB for the MCL-D, both of which were well below the 5-dB difference commonly considered to be needed to interpret audiometric differences as clinically significant. Overall, we conclude that the estimates of the MCL-A and MCL-D obtained here with spondees singlets and spondees triplets were both comparable to one another and to the estimates obtained with connected discourse.

Previous work supports the view that listeners may identify a range of levels as comfortable when listening to speech (Berger & Lowry, 1971; Pollack, 1952; Punch, Joseph, et al., 2004; Punch, Rakerd, et al., 2004). The MCL-A and MCL-D results from this study provide an example of this. On each run, a participant gave a loudness rating of 4 (comfortable) to both the MCL-A and MCL-D. Despite this perceived equivalence, the measured stimulus intensity for the MCL-D exceeded the measured intensity for the MCL-A on 96% of all runs in this study.

The MCL-A and MCL-D results in Figure 3 offer a basis for estimating the expected size of the speech comfort level range for listeners with normal hearing. For the 36 listeners in the age range of 18–29 years, the level difference between the MCL-D and MCL-A averaged 10.4 dB (SD = 3.2 dB). That difference was statistically significant, \( t(35) = 19.60, p < .001 \), and it represents a best estimate of the MCL range for this group. A confidence interval computed from these results indicates that there is a 95% probability that the MCL range for any young adult with normal hearing should fall between 9.3 and 11.5 dB.

MCL and UCL

Figure 4 shows the MCL (average of the MCL-A and MCL-D) and UCL results for the experiment. The primary purpose of this study was to determine whether the MCLs and UCLs obtained with spondees, presented as either singlets or triplets, would be comparable with the MCLs and UCLs obtained with connected discourse. The results shown in the figure clearly indicate that they were comparable. The mean MCLs and UCLs for the spondees group differed from those for the discourse group by less than 3 dB (MCL difference = 2.7 dB, UCL difference = 1.9 dB), and the means for the triplets group differed from those for the discourse group by less than 2 dB (MCL difference = 1.2 dB, UCL difference = 0.9 dB). Statistically, there were no significant differences among the three stimulus groups for either the MCL (\( F = 0.739, p = .485 \)) or the UCL (\( F = 0.089, p = .915 \)), and stimulus type effect sizes were small to medium (\( \eta^2 = .043 \) for the MCL and \( .005 \) for the UCL). We conclude that spondees can be used as a valid alternative to connected discourse when measuring both the MCL and UCL for speech.

\(^4\) Other effect size estimates for stimulus type found in this study were \( \eta^2 = .043 \) for the MCL and \( \eta^2 = .005 \) for the UCL. On average, stimulus type was found to account for about 3% of the variance in the various quantities measured (mean \( \eta^2 = .033 \)). Cohen (1988) provided the following effect size benchmarks for eta square: small (\( \eta^2 = .01 \)), medium (\( \eta^2 = .06 \)), and large (\( \eta^2 = .14 \)).

Figure 3. Mean MCL-As and MCL-Ds for the three experimental groups, with standard deviations. MCL = most comfortable level (A = ascending; D = descending).
Reliability

In this section, we report on the reliability of the MCL and UCL test results.

Intrasession Reliability of the MCL and UCL

To assess intrasession reliability on both Days 1 and 2, we calculated the percentage of the 36 participants who (a) had retest measurements of the MCL or UCL that exactly matched their corresponding test measurements, (b) had retest measurements that were within 5 dB of the corresponding test measurements, and (c) had retest measurements that were within 10 dB of the corresponding test measurements.

For the MCL test and retest, levels exactly matched for 19% of the participants, were within 5 dB for 75% of the participants (including the 19% who had exactly matched), and were within 10 dB for 97% of the participants. On average, the MCLs measured on retests differed from the corresponding tests by about 1 dB (+0.6 dB on Day 1 and −1.1 dB on Day 2).

For the UCL, nearly half of all participants had matching test and retest scores (45.6% for Days 1 and 2 combined), 89% had scores within 5 dB, and 99% had scores within 10 dB. There was a tendency for the UCL retest results to be slightly higher than the corresponding test results on both days. On average, they were 2.2 dB higher on Day 1 and 1.4 dB higher on Day 2.

Intersession Reliability of the MCL and UCL

Intersession reliability results were essentially the same for the measurements of the MCL and UCL. Eleven percent of the 36 participants exhibited levels on Day 2 that exactly matched their Day 1 levels, 72% had Day 2 levels within 5 dB of their Day 1 levels, and 97% had Day 2 levels within 10 dB of those on Day 1. The average difference between the Days 2 and 1 MCL values was +0.6 dB, and for the UCL measurement, it was +1.8 dB.

Test Times

To determine the degree to which the test protocol used in this study can be considered clinically efficient, we used digitally stored data to track the amount of time needed to complete each cycle of testing and retesting. Table 1 reports those times (means and standard deviations, in minutes) for familiarization, SRT, and MCL/UCL conditions, for each of the three experimental groups.

At a glance, two findings stand out. First, familiarization took longer to complete than any of the SRT measurements, and the MCL/UCL measurements took longer to complete than any of the SRT measurements. With respect to efficiency, the latter is not surprising, given that the MCL/UCL measurements consisted of quantifying three separate measures: MCL-A, MCL-D, and UCL. In fact, the longest time needed to complete the MCL/UCL measurements (3.43 min) was only 2.12 min longer than the shortest time to complete the SRT measurements (1.31 min), which reflects well on predicted clinical efficiency.

To predict the time needed to complete each of these tests in a typical clinical setting, mean times and their variability were estimated by combining the times needed for spondee familiarization during the test session on Day 1 and doubling the times needed to obtain the SRT, MCL, and UCL during the initial test session on that same day. Given that we tested our listeners in their preferred ear only, doubling of these latter values was done to arrive at a realistic estimate of the time needed to test both ears of a clinical patient. The results are given in Table 2, which shows that the predicted total (mean) times were 10.7, 12.3, and 12.4 min, respectively, for spondee singlets, triples, and connected discourse. It should be noted that these times do not include instructions. Our informal assessment of the time needed to read the relevant instructions aloud indicated that approximately 2 min should be added to the times given in Table 2 to estimate the overall time needed to complete familiarization and all the SRT, MCL-A, MCL-D (and consequent MCL), and UCL measurements in both ears of a clinical patient.

Clinical Considerations

The protocol described in this study is intended for use with adults and older children who can follow the instructions and execute them with accuracy and precision. If followed, the recommended protocol promises to provide clinically useful information in a reliable, valid, and efficient manner that embodies best practices in speech audiometry.

Here, as well as in earlier research by others and ourselves, an emphasis has been on the importance of separating measurements of the MCL and UCL by a time interval sufficient to preclude or reduce the influence of exposure to higher speech intensities on exposure to lower intensities.
Specifically, these studies caution against measuring the UCL before the MCL and against establishing an MCL-D before establishing an MCL-A, as either technique could inadvertently raise the MCL. That conclusion has been based largely on experimental measurements in a single ear of listeners. Based on tests performed in the preferred ear of young adults with normal hearing, Punch, Rakerd, et al. (2004) found that the average MCL of listeners with normal hearing was almost 7 dB higher when it followed a UCL test by about a minute, and the difference increased substantially when the time between the UCL and MCL was reduced even further. That outcome suggests that listeners base their speech loudness judgments on reference levels that change with their moment-to-moment auditory experience.

Use of the protocol described in this study requires clinicians to abandon the long-held notion that each separate test—the SRT, MCL, and UCL—should be administered procedurally in one ear (usually the better ear) and then in the opposite ear. Instead, the entire test sequence, consisting of the SRT, MCL-A, MCL-D, and UCL, should be administered in one ear before being administered in the opposite ear. That accommodation is essential for the efficient implementation of the recommended protocol, which is aimed at simplifying the integration of four separate speech audiometric tests in each ear.

With respect to how the proposed protocol might affect the measurement of the MCL in a second ear after an initial test of the UCL in the opposite ear, the available evidence suggests that its clinical impact can be expected to be minimal, for several reasons. First, even the brief, natural time delay that occurs when switching between ears can reduce any negative influence of a prior (opposite-ear) UCL test, and time devoted to other intervening steps could reduce it further. For example, the clinician might inform the patient that the same tests will next be administered in the opposite ear and offer a reminder of the specific order of those tests. Second, and more importantly, the fact that the SRT—with its inherently lower levels—is the first test administered in the second ear will have the effect of lowering the listener's loudness reference for subsequent testing. Third, shifting the testing to the second ear can be expected to result in an additional shift in the listener's loudness reference by virtue of stimulating different auditory pathways. Such a shift in reference is especially apt to occur in cases of asymmetrical hearing loss. Although these expectations have not been verified by formal experimentation, they have been supported by preliminary observations based on pilot testing in our laboratory. Finally, as suggested by Punch, Rakerd, et al. (2004), a post hoc correction might be added if these combined factors were to prove insufficient to eliminate all influence of establishing the UCL in one ear on the MCL in a subsequently tested opposite ear.

### Follow-Up Experiment

The mean intensity levels for the MCL (loudness rating of 4) and UCL (rating of 6) shown in Figure 4 were both near the lower limit of what Cox et al. (1997) described as the typical normal range of results for the Contour Test.

<table>
<thead>
<tr>
<th>Table 1. Time (with standard deviations), in minutes, for each of the three stimulus groups to complete each component of the experiment.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Time (with standard deviations), in minutes, predicted from the experimental data to be needed for completion of a clinical session that includes familiarization, speech recognition threshold testing in both ears, and most comfortable level/uncomfortable level testing in both ears.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note. Fam = familiarization; SRT = speech recognition threshold; MCL = most comfortable level; UCL = uncomfortable level; T1 = test, Day 1; T2 = test, Day 2; R1 = retest, Day 1; R2 = retest, Day 2.
Given this unexpected finding, we decided to perform a follow-up experiment.

We conjectured that this result might be explained, in part, by the age range of our participants, as they were all young adults with normal hearing ranging in age from 18 to 29 years. From their adolescence to the present day, that age group has received targeted messaging about the need to protect their hearing by limiting the volume of music and other audio signals played through their personal listening devices (e.g., Punch, Elfenbein, & James, 2011). We speculated that they may have instilled these cautionary messages to a greater extent than older generations and, in turn, may have responded more conservatively when making loudness ratings.

To examine this possibility, we recruited a separate group of 12 participants who were 30–39 years old. Like the younger listeners recruited for the main experiment, individuals in this decade-older group were accomplished speakers of English and exhibited normal hearing. All of these participants passed a hearing screening in both ears at 500, 1000, 2000, and 4000 Hz, and they all had SRTs between 0 and 10 dB HL (M = 5.8 dB). We ran a 1-day test and retest on this older participant group that followed the same protocol described above for Day 1 testing of the younger participants. All of the older adults were tested only under the spondee singlets condition. Their MCL and UCL results, averaged over test and retest, together with those of a comparison group of participants from the main experiment who were tested under the same conditions (spondee singlets, Day 1 only), are given in Figure 5.

The results show an age trend that is consistent with our conjecture, with the MCLs 5.9 dB higher and the UCLs 6.7 dB higher, on average, for the older group. The age factor was statistically significant for the UCL, \( t(22) = 2.28, p < .033 \), and marginally significant for the MCL, \( t(22) = 2.00, p < .058 \), and it had a large effect size for both the MCL (Cohen’s \( d = 0.82 \)) and UCL (\( d = 0.93 \)). We characterize this outcome as tentative evidence that the current generation of younger adults with normal hearing prefer to listen to speech at somewhat lower levels than do their older counterparts, very possibly owing to a heightened awareness of the dangers of exposure to louder sounds.

**Conclusions**

In this study, we reached the following conclusions:

1. Spondee words offer a valid alternative to connected discourse when assessing a listener's MCL and UCL for speech.
2. Single spondees provide estimates of the MCL and UCL that are at least as reliable and valid as spondee triplets and are recommended over triplets to replace connected discourse for these measurements.
3. With respect to efficiency, the use of singlets, compared with both triplets and discourse, required less time to complete measurements of both the MCL and UCL, as would occur in a typical clinical protocol. Overall, therefore, we can reasonably predict that the use of singlets will lead to optimal reliability, validity, and clinical efficiency when making these measurements.
4. The 18- to 29-year-old participants in our main experiment, when compared with the 30- to 39-year-old participants in the follow-up experiment, exhibited notably lower MCLs and UCLs. Their mean levels were near the lower limit of the typical normal range of results for the Contour Test (Cox et al., 1997). This led us to speculate that younger populations may be responding with relative caution due to repeated public messaging urging younger age groups to avoid loud music and other sounds. Additional research will be needed to validate such a speculation.
5. Although the test protocol described here can be performed manually, computer automation is recommended to optimize the protocol for general clinical use. The current trend to incorporate microprocessor-based test procedures into contemporary audiometric test procedures seems to make automation of the current protocol, or a similar protocol, highly feasible. Although not done here, it might be possible to combine and shorten the instructions for familiarization and SRT testing with those for MCL and UCL testing to minimize clinical test time. In an automated protocol, the instructions could easily be displayed on a computer interface, as opposed to a printout. Other modifications that further minimize clinical test time without sacrificing the reliability or validity of speech audiometric measures are worthy of additional investigation. Based on the values obtained in the protocol proposed here, automation might be extended.
further by implementing WRS testing at a predetermined level (using either the obtained MCL or a sensation level setting, referenced to the SRT), determining the need for contralateral masking during SRT and loudness testing, and, if needed, establishing the appropriate masking levels.

6. Our two earlier articles on the topics of the MCL and UCL established that, because the MCL is a range of levels and not a fixed level, one goal in speech audiometry should be to quantify that range. Results of this study indicate that there is a 95% probability that the MCL range for any young adult with normal hearing will fall between 9.3 and 11.5 dB. That range of approximately 10 dB may be viewed as the range of comfortable speech for individual listeners with normal hearing, and it is the key concept underlying our recommendation to incorporate the test procedures evaluated in this study. An average MCL can be determined from the range, and the range itself could be useful as a refinement to prescriptive hearing aid fitting formulas aimed at maintaining comfortably loud speech for individual listeners with hearing impairment.

Acknowledgments

The second author presented preliminary data from this study as a poster at the November 2017 annual convention of the American Speech-Language-Hearing Association held in Los Angeles, CA. The first author presented a fuller account of the data in a poster at the April 2018 meeting of the American Academy of Audiology held in Nashville, TN. The authors wish to thank two anonymous reviewers for their very helpful comments on an earlier version of the manuscript.

References


Appendix A

Familiarization Instructions (All Stimulus Groups)

This study will involve several steps, and I will be giving you specific instructions for each step. The first step involves familiarizing you with the two-syllable words that will be used in the experiment. Examples of such words are baseball and railroad. All the words will appear in alphabetical order on the computer screen, and you should use the mouse to click on the word you hear. Printout 1 (on the wall) illustrates what you will see on the screen. For familiarization purposes, you’ll hear all the words at the same volume setting. Just follow the visual cue “Press here to start” on the screen when you’re ready to begin. The ear icon on the screen is your cue to listen carefully. You shouldn’t talk during the test, but if you need to ask a question, I’ll be able to hear you. When you’re finished with this part, I’ll give you further instructions.

Appendix B

SRT Test Instructions (All Stimulus Groups)

Next, I want to find the softest speech that you can hear. The words you just heard will be presented randomly, one at a time, and your job is to identify each word. Again, click on “Press here to start” to begin the test, and listen carefully when you see the ear icon. The words will become softer and softer. Click on each word you hear. Make an effort to respond to each word, no matter how soft, even if you have to guess. Click on “Could not tell” whenever you are too uncertain to guess.

Appendix C

MCL/UCL Test Instructions (Spondee Singlets Group Only)

The purpose of the next test is to obtain your judgments of how comfortable words are when heard at different volume settings. Again, you’ll hear the two-syllable words, one at a time, at a volume that gets progressively louder or softer. Here, your job is not to identify each word but to rate the loudness of each word, using the mouse button to click on a number from 1 to 7 that corresponds to its loudness.

Some of the speech you’ll be hearing in this test will be louder than the speech you heard before. For your ratings of 6 and 7, be as tolerant as possible of these louder sounds. In making your loudness judgments, pretend that you’re listening to the voice of a radio broadcaster. Select a rating of 6 to indicate that the broadcast is louder than you would normally choose but still OK and a rating of 7 to indicate that the broadcast is louder than you would ever choose. You should not be concerned that any of these sounds you hear will be loud enough, or long enough, to affect your hearing.

Printout 2 (on the wall) illustrates what you’ll see on the computer screen. On the 7-point scale, 1 is very soft and 7 is uncomfortably loud. Descriptions of the different loudness levels are given below each number, and you should click on the number, not its description. Twice during this test, there will be a brief pause.